PB86-207511

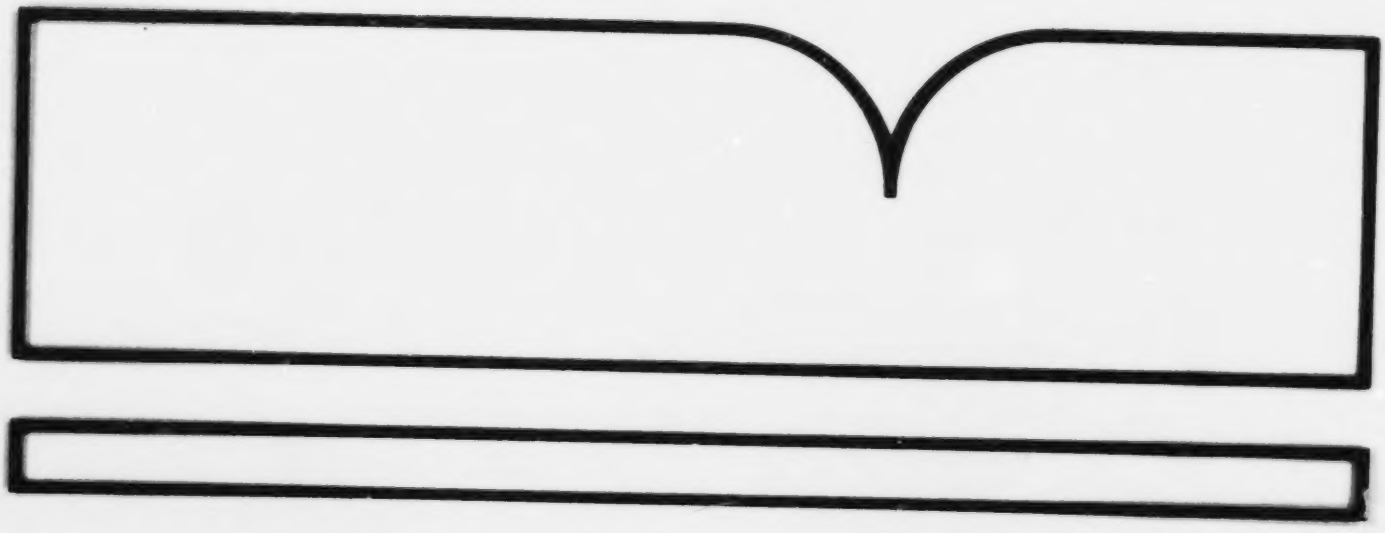
Structure-Activity Relationships (SARs) among Mutagens and Carcinogens: A Review

Case Western Reserve Univ., Cleveland, OH

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1986



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completir

1. REPORT NO.	2.	3. R
EPA/600/J-86/090		PB86-207511

4. TITLE AND SUBTITLE
Structure-Activity Relationships (SARs) Among Mutagens and Carcinogens: A Review (Journal Version)

5. REPORT DATE
1986

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
M.R. Frierson, G. Klopman, and H.S. Rosenkranz

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Chemistry
Center for the Environmental Health Sciences
Case Western Reserve University
Cleveland, Ohio

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
CR812264-01

12. SPONSORING AGENCY NAME AND ADDRESS
Health Effects Research Laboratory - RTP, NC
Office of Research and Development
US Environmental Protection Agency
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE
EPA/600/11

15. SUPPLEMENTARY NOTES
Journal - Environmental Mutagenesis 8: 283-327 (1986) March Issue #2

16. ABSTRACT

This review is an introduction to methods for evaluating structure-activity relationships (SARs), and, in particular, to those methods that have been applied to study mutagenicity and carcinogenicity. A brief history and some background material on the earliest attempts to correlate molecular structure and biological activity are included. Most of the discussion focuses on modern methods utilizing extrathermodynamic and physical property variables such as the Hansch method and SIMCA, and approaches based on molecular connectivity such as the ADAPT, CASE, and Enslein methods. In general, the latter class is potentially the most useful in the study of the large and structurally diverse databases so often encountered in the study of mutagenicity and carcinogenicity. They also are not very sensitive to lab-to-lab variances in reported activities and outright misclassifications in activities of some compounds. This is chiefly because the statistical treatments used in these methods tend to dilute the importance of outliers. The methods using physicochemical and extrathermodynamic variables are especially important in relatively small, congeneric databases and can help fine-tune the role of physicochemical properties in mechanistic hypotheses. All of the above methods have been used to look at mutagenicity and carcinogenicity and some of the results reported in the literature are reviewed here. As far as specific methods go, ADAPT, CASE, SIMCA and the Enslein approach all seem to have similar classification powers (in the range of 75-95%), depending very much on the database studied. The emphasis in this review is on showing that the use of these computer-aided storage, retrieval and analysis techniques is a timely approach to predicting and even understanding the toxicity of environmental substances. However, each of the methods discussed is still under development, and their potential usefulness for predictive purposes is still being explored.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	46
	20. SECURITY CLASS (This page)	22. PRICE
	UNCLASSIFIED	

EPA/600/J-86/090
JOURNAL ARTICLE

Environmental Mutagenesis 8:283–327 (1986)

Structure-Activity Relationships (SARs) Among Mutagens and Carcinogens: A Review

Manton R. Frierson, Gilles Klopman, and Herbert S. Rosenkranz

Department of Chemistry (M.R.F., G.K.) and Center for the Environmental Health Sciences, School of Medicine (H.S.R.), Case Western Reserve University, Cleveland

This review is an introduction to methods for evaluating structure-activity relationships (SARs), and, in particular, to those methods that have been applied to study mutagenicity and carcinogenicity. A brief history and some background material on the earliest attempts to correlate molecular structure and biological activity are included. Most of the discussion focuses on modern methods utilizing extrathermodynamic and physical property variables such as the Hansch method and SIMCA, and approaches based on molecular connectivity such as the ADAPT, CASE, and Enslein methods. In general, the latter class is potentially the most useful in the study of the large and structurally diverse databases so often encountered in the study of mutagenicity and carcinogenicity. They also are not very sensitive to lab-to-lab variances in reported activities and outright misclassifications in activities of some compounds. This is chiefly because the statistical treatments used in these methods tend to dilute the importance of outliers. The methods using physicochemical and extrathermodynamic variables are especially important in relatively small, congeneric databases and can help fine-tune the role of physicochemical properties in mechanistic hypotheses. All of the above methods have been used to look at mutagenicity and carcinogenicity and some of the results reported in the literature are reviewed here. As far as specific methods go, ADAPT, CASE, SIMCA and the Enslein approach all seem to have similar classification powers (in the range of 75–95%), depending very much on the database studied. The emphasis in this review is on showing that the use of these computer-aided storage, retrieval and analysis techniques is a timely approach to predicting and even understanding the toxicity of environmental substances. However, each of the methods discussed is still under development, and their potential usefulness for predictive purposes is still being explored.

Key words: mutagenicity, carcinogenicity, artificial intelligence, ADAPT, CASE, SIMCA, QSAR, SAR

INTRODUCTION

Structure-activity relationships (SARs), considered here as those techniques that treat the relationship between biological activity and molecular structure in a statistical

Received August 7, 1985; revised and accepted November 11, 1985.

Address reprint requests to Gilles Klopman, Department of Chemistry, Case Western Reserve University, Cleveland, OH 44106.

© 1986 Alan R. Liss, Inc.

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

or quantitative manner, have been found to be extremely useful in the field of drug design [Martin, 1978, 1981]. They have only recently come to be used in the field of toxicology and, explicitly, mutagenicity and carcinogenicity. There are probably many reasons for this. Among them may be the fact that those most interested in mutagenicity have had little exposure to the methods, which started developing largely among medicinal chemists in the late 1960s and continue to this day (new approaches to SARs are still being developed). This review is an attempt to bridge these communication and knowledge gaps.

Another possible reason for the lack of use of SAR methodologies in genetic toxicology relates to the fact that the application of SARs to this field presents some unique problems that are less likely to be encountered in the field of drug evaluation and design. One important difference is that mutagenicity, being the result of cell penetration, bioactivation, interaction, and modification of DNA, and various error-free and error-prone DNA repair processes, is not the well defined biological activity that, say, anticholinesterase activity is. This is reflected in the great diversity of chemical structures and functionalities that have been found to be associated with mutagenicity. Related to this problem are the plethora of tests that are believed to have a relationship to mutagenicity and genotoxicity. It is not even certain that the tests have a qualitative relationship to one another much less a quantitative one [Pet-Edwards et al, 1985a,b]. However, the rewards for obtaining a molecular structural interpretation of genotoxic effects are great, not just for purposes of classification and prediction but also for understanding the fundamental processes of genotoxicity in all of its manifestations. In the next few sections, the background and methods of the best known and most often used methods will be discussed. In no case will a rigorous introduction be given; rather, the vocabulary and some crucial references to more detailed reviews of specific methods are provided. Additionally, when those methods are known to have been applied to the study of mutagenicity and carcinogenicity, these reports are discussed. The methods are divided into two broad and somewhat arbitrary categories based on the type of variables they utilize: 1) extrathermodynamic approaches and 2) "connectivity" approaches. The extrathermodynamic approaches emphasize the physical properties of the molecules under study and are largely limited to SAR studies among congeneric compounds.

The connectivity approaches emphasize the atomic construction of the molecules, including but not limited to bond connections. Many of the connectivity methods, while handling databases that are noncongeneric (ie, heterogeneous), are not quantitative. That is, they would not be able to predict an expected potency (eg, revertants/nanomole) but only whether a tested compound belonged to a particular activity class or not. The approaches clearly are not necessarily mutually exclusive, and attempts have been made to combine them. Another important part of the classification is the method of statistical treatment of the data.

EXTRATHERMODYNAMIC APPROACHES

Roots in Physical Organic Chemistry

The extrathermodynamic approach to structure-activity correlations actually has its roots in physical organic chemistry, in which the activity to be correlated was chemical reactivity. L.P. Hammett [1940] was able to relate the ionization constants of a series of substituted phenylacetic acids (1; see Fig. 1) to a series of substituted benzoic acids (2) by the following relation:

$$\log(K_X/K_H)_{pa} = 0.49\log(K_X/K_H)_{ba}. \quad (1)$$

where K represents the substituted (X) and unsubstituted (H) ionization constants of phenylacetic acids (pa) and benzoic acids (ba). (We follow here the Hirsch [1974] demonstration of Hammett's equation.) If we assume that the only effect the substituents have on the acidity of these compounds is electronic (a reasonable assumption in that they are located some distance from the reactive center), the relation makes perfect sense, because one would expect the electronic effect of the substituents, X, to be qualitatively the same in each series, attenuated only by the intervening methylene group. Then, because the ionization constants, K, can be related to free energies as:

$$\Delta G = -RT \ln K, \quad (2)$$

the above equation, 1, can be rewritten:

$$(\Delta G_X - \Delta G_H)_{pa} = C_o(\Delta G_X - \Delta G_H)_{ba}. \quad (3)$$

This equation defines a simple, linear free-energy relationship, albeit in a particularly awkward form. Taking the benzoic acid series as a standard reaction set, we can relate all other reactions to this one. Then the quantities associated with this series become constants, such as:

$$\log(K_X/K_H)_{ba} = \sigma_X. \quad (4)$$

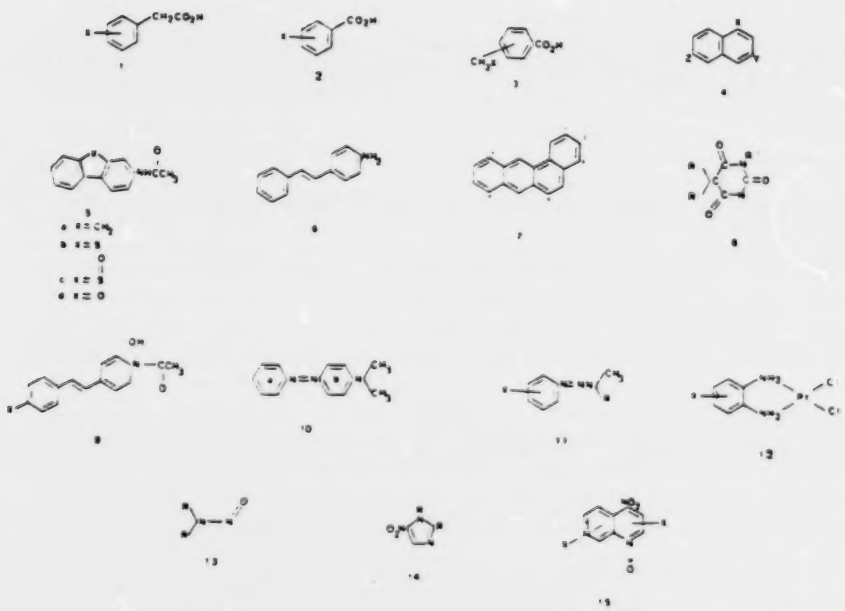


Fig. 1. Molecular structures chart.

This quantity, σ_X, is called a *substituent constant* and represents the electronic effect (relative to H) of the substituent X in the standard reaction. It turns out that the effect is different depending on where the substituent is attached. Since the effect of a substituent located *ortho* may be steric as well as electronic (because of its proximity to the reaction center), Hammett defined only two substituent constants per substituent, σ_m for *meta* substituents and σ_p for *para* substituents. The slope in equation 1 represents the sensitivity of the new reaction series to the electronic effects of the σ's relative to that of the standard reaction (whose slope is 1). It is given the symbol ρ. With these symbols, the Hammett equation for equilibria, in its standard notation, is written as:

$$\log(K_X/K_H)_{\text{new series}} = \rho_{\text{new series}} \cdot \sigma_{ba} \tag{5}$$

A Hammett-type equation can also be written for reaction rates based on the Bell-Evans-Polanyi principle [Dewar and Dougherty, 1975]. This principle relates activation energies, ΔG, to the thermodynamic quantities $\Delta G°$:

$$\Delta G = \Delta G° + C. \tag{6}$$

The Hammett equation for kinetic phenomena is written analogously to the equation for equilibria by replacing the equilibrium constants by kinetic rate constants, k_i:

$$\log(k_X/k_H)_{\text{new series}} = \rho_{\text{new series}} \cdot \sigma_{ba} \tag{7}$$

Since the original inception of Hammett's σ_m and σ_p constants, a variety of other constants have been defined and used successfully to model the effect of changes in molecular structure on many kinds of chemical reactions and properties [Hirsch, 1974; Lowry and Richardson, 1976]. They all have in common an ultimate relation to thermodynamic quantities even though they cannot be derived directly from the postulates of formal thermodynamics. For this reason, they have been termed *extrathermodynamic quantities*.

Early Attempts to Correlate Biological Activity to Pnysical Properties: Brief Hisiory

The first attempts to correlate a biological activity to molecular properties occurred at the turn of the century. Meyer [1899] and Overton [1901] both looked to oil/water partition coefficients to explain the potencies of certain nervous system drugs [Osman et al, 1979]. The first really quantitative approach to the treatment of biological activities and molecular structure was that by Collander [1954], who showed that the rate of transport through cellular material was proportional to the log of P (the partition coefficient between an organic solvent and water). This was extended by Hansch et al [1963] and Hansch and Fujita [1964], who used the Hammett approach to relate biological response rate constants to σs and $\log(P_X/P_H)$s. Hansch defined $\log(P_X/P_H)$ as π and obtained the following general equation:

$$\log(1/C) = -k \cdot \pi_x^2 + k' \cdot \pi_x + \rho \cdot \sigma_x + k'' \,. \tag{8}$$

where C is the molar concentration required for a particular biological activity, ρ and σ_x have the meanings given above, and the ks are constants. Since this beginning, a number of other parameters (often called *descriptors*) have been considered and used.

However, for most purposes the following regression equation has become the standard of the Hansch method:

$$\log(1/C) = a + b\cdot\pi_x + c\cdot\pi_x^2 + d\cdot E_s + e\cdot MR + f\cdot\sigma_x, \qquad (9)$$

where a–f are coefficients. The πs and σ are the same as above. The new regression descriptors are E_s, which is a parameter defining the steric properties of a substituent, and MR, which defines its polarizability (vide infra). Clearly, since the substituent parameters are defined with respect to hydrogen, they can be used only for congeneric databases. As will be shown below, global properties, such as the log P, may be used as a parameter in heterogeneous databases. For a complete and thorough introduction to the Hansch approach, there are a number of reviews [Osman et al, 1979; Hansch, 1978].

Hansch Method

Hydrophobicity parameters. The quantity P, the partition coefficient, usually octanol-water, is presumed to model the partitioning of a molecule with the various biophases found in the intact, living organism. Other organic solvent systems have been tried, but the octanol-water system has been found to be adequate for most cases. Hydrocarbons, in particular, were found to give poor correlations in binding studies [Osman et al, 1979]. This is thought to be due to water-solute desolvation on passage from the aqueous to the hydrocarbon organic phase [Osman et al, 1979]. The correlation of such properties with the octanol-water system is better because the octanol contains significant amounts of water [Osman et al, 1979].

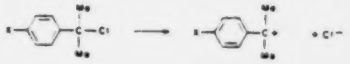
The rationale behind the inclusion of the square of the solubility properties is interesting in what it has to say about the relation between transport and interaction properties of biologically active molecules. Hansch [1971] suggested that, in order to reach the high concentrations of the solute at the active site necessary for biological action, a number of partitions between hydrophobic and hydrophilic phases must occur. This means that highly hydrophobic species will tend to collect in the hydrophobic phases, which will actually inhibit efficient transport, and an extremely hydrophilic species will tend to concentrate solute species in the hydrophilic phases, again to the detriment of efficient transport. The optimal value of the partition coefficient will then clearly lie somewhere between the two extremes. Hansch suggested the parabolic approximation, although there are other approximations for this nonlinear, optimal solubility behavior [Flynn and Yalkowsky, 1972; Yalkowsky and Flynn, 1973; Higuchi and Davis, 1970; Kubinyi, 1977].

There do exist tabulations of P and logP values [Leo et al, 1971; Leo, 1985]. In addition, there are routines available for calculating P and logP empirically [Leo et al, 1971; Chou and Jurs, 1979a; Hopfinger and Battershell, 1976; Rekker, 1977; Rekker and deKort, 1979; Meyer et al, 1982; Klopman and Iroff, 1981; Klopman and Namboodiri, 1985].

Electronic parameters: σ, σ°, σ_p^-, σ_p^+, σ_I, σ_R°, and σ^*. The σ terms of the Hansch equation are meant to represent the electronic effects of a substitutent. There are a number of problems with the σ as originally formulated. For instance, the σ_p can interact with some reactive centers by "through-resonance" [Clark and Perrin, 1964] (see Fig. 2). One approach for dealing with this problem was reviewed by H.C. Brown and L.M. Stock [1963]. There are two different standard reactions used, one for dealing with the case of an electron-donating para-substituent to an electron-deficient reaction center (equation 10) and one for dealing with an electron-withdraw-

Fig. 2. An example of through-resonance.

ing substituent with an electron-rich reaction center (equation 11). The resulting σs are σ_p^+ and σ_p^-, respectively.



(10)

(11)

The σ_ms were assumed to be the same as the Hammett σ_ms. This approach was also criticized on the grounds that the σs were still strongly dependent on the mode of interaction [Wepster et al, 1959, 1973; Taft and Lewis, 1958, 1959a,b]. Taft [1960] and Taft and associates [1959] suggested using species such as 3 (see Fig. 1) for the standard reaction, in which an intervening methylene group would prevent the through-resonance. This is the basis of the σ° constant. Even this constant can be broken up into two contributing quantities, σ_I, for the inductive, through-bond contribution and σ_R°, the resonance contribution to the electronic effects:

$$\sigma^\circ = \sigma_I + \sigma_R^\circ. \tag{12}$$

The σ_I is defined based on still another σ from aliphatic systems (σ^*):

$$\sigma_I = 0.54\sigma^*. \tag{13}$$

This σ was devised to deal with electronic effects in aliphatic systems by Taft [1952, 1953] at the suggestion of Ingold [1930]. Because electronic effects acting through localized (ie, aliphatic) bonds generally do not act over great distances, the problem of trying to evaluate electronic effects in these systems was complicated by the probability of interference from steric effects. Taft solved this problem by noting that the acid hydrolysis of benzoic acid esters was insensitive to electronic effects (ie, these reactions had a very small ρ value). On the other hand, the base-catalyzed hydrolysis showed a strong dependence on electronic effects (these reactions had large values of ρ). Since these two reactions have very similar transition states, it might be possible to separate the electronic effects from the steric ones (Fig. 3). Taft reasoned that the rate effects observed in the acid hydrolysis of *aliphatic* acid esters on substitution (equation 14) were due to steric effects. A Hammett-type relation defining a new substituent constant, E_s, was suggested that would account for these steric effects on rates of hydrolysis:

$$\log(k_X/k_H) = E_s, \tag{14}$$

where k_X is the reaction rate of the substituted ester and k_H is the reaction rate of the unsubstituted ester. Then, armed with the knowledge that the steric effects are essentially the same in each reaction, he examined the same aliphatic series under

basic catalysis to obtain a σ supposedly free of steric effects and representing a "true" inductive effect (resonance effects were not possible in these aliphatic systems). This is σ^*:

$$\sigma^* = \frac{\log(k_X/k_H)_B - \log(k_X/k_H)_A}{2.48} \tag{15}$$

One other approach at generality has been suggested based on statistical techniques that would allow one to obtain the "best fit" σ values from a variety of standard reactions. It is based on the following equation:

$$\log K_{ik} = \phi_k + \rho_k \cdot \sigma_i + e_{ik}, \tag{16}$$

where the ks refer to the reaction standard, i refers to substituents and K is the equilibrium (or rate constant), ϕ_k is the log of the unsubstituted rate or equilibrium constant, and e is the experimental and model error [Sjostrom and Wold, 1974, 1976]. The quantities are minimized to obtain the best fit values.

Finally, an attempt was made to tie the σ to quantities that could actually be calculated from first principles without direct recourse to experimental measurements [Dewar and Grisdale, 1962]. Dewar and Grisdale [1962] divided the σ constants into two components, electrostatic and resonance with induction; ie, the σ were then represented as:

$$\sigma_{ij} = F/r_{ij} + Mq_{ij}, \tag{17}$$

or

$$\sigma_{ij} = F'/r_{ij} + M' \cdot \pi_{ij}. \tag{17}$$

where F(F') is the field exerted by a substituent attached at atom i and r_{ij} is the distance between atom i and the reactive functionality attached at atom j, M(M') is the combined π-resonance inductive effect of the substituent, q_{ij} is the charge overlap between i and j, and π_{ij} is the π-charge overlap. This particular system has not found much use in QSAR but with development could become more useful [Osman et al, 1979].

The evolution of σ constants is complex and, at times, confusing because of the many similar notations, each with a slightly different connotation. In Table I we have summarized the various σ as discussed here, giving a brief definition. Even more distressing is the problem one has in making a choice of a scale when attempting a

Fig. 3. Transition states for base-catalyzed and acid-catalyzed ester hydrolysis.

TABLE I. Sigma Constants and Their Meanings

Constant	Meaning
σ_m, σ_p	The original Hammett constants
σ_p^+, σ_p^-	Constants explicitly for use with through-resonance
σ°	The Taft constant with through-resonance blocked
σ_I	Inductive electronic contribution to in aromatic systems (see eq 12)
σ_R°	Resonance electronic contribution to σ° (see eq 12)
σ^*	Inductive electronic contribution of aliphatic systems

correlation. Practitioners of QSAR suggest use of the σ° or σ_I scales for general applications when not much is known about the underlying molecular mechanisms of activity [Osman et al, 1979].

Steric parameters

Taft steric parameters. The Taft steric substituent constant is of interest in its own right and has been useful in biological QSAR [Osman et al, 1979; Hansch, 1978]. In fact, the problems associated with the three-dimensional character of molecules is one of the more underdeveloped areas of QSAR work. The Taft steric parameter, E_s, is still probably the most widely used geometrically linked parameter in Hansch analyses, and Hansch et al [1973] have published a review of available E_s values. Another compilation of steric constants (and other parameters) is that of Exxner [1978]. There are a number of attempts to overcome some of the shortcomings of the Taft parameters [Simon et al, 1977; Verloop et al, 1976; Hopfinger and Potenzone, 1982; Hansch et al, 1973].

Minimal topological differences (MTDs). An interesting approach to the problem of steric effects is the minimal topological difference (MTD) or, as Osman et al [1979] call it, minimal steric difference (MSD). This technique of Simon et al [1977] looks at "volume" differences between an idealized archetypal structure and a test structure. There is no attempt actually to calculate molecular volumes for comparisons. Instead, a procedure has been devised that calculates, in a relative way, nonoverlapping volumes. The procedure works as follows (we use here the description and notation of Niculescu-Duvaz et al [1981]): 1) Given N molecules of some specific biological activity, one superimposes them, atom by atom and excluding hydrogen atoms, to obtain a framework of n atomic vertices. This framework is called a *hypermolecule*. 2) Then, for each molecule, M_i, an associated vector is calculated by summing components x_{ij}, obtained by assigning the values of 1.0, 1.5, and 2.0 to each atom of M_i coincident with a vertex j in the hypermolecule and belonging to the first, second, and third periods of the periodic table, respectively. Those vertices for which M_i has no coincident atoms are assigned an x_{ij} of 0. 3) An initial guess is made as to the shape of the biological receptor by choosing a standard structure, S, derived from consideration of the structures and activities of the set of N molecules. It is represented by another n-dimensional vector with components ϵ_j assigned values of 1, −1, and 0 for vertices that belong to the receptor wall, the receptor cavity, and sterically unimportant regions, respectively. 4) Then the MTD_i is calculated for each molecule M_i as:

$$MTD_i = s_i + \Sigma\epsilon_j \cdot x_{ij}, \tag{18}$$

where s_i is the number of vertices in M_i with $\epsilon_j = -1$. The standard structure, S, is refined by optimizing the equation below for biological activity with respect to the ϵs:

$$BA_i = a - b \cdot MTD_i. \tag{19}$$

The resulting optimized standard structure, S*, is the best complementary fit to the receptor site, and all succeeding MTD_i are calculated with respect to it. The MTD_i are then often used in the Hansch equation in place of E_s. This method does not appear to work well for a set of N compounds that were chemically diverse; then, the choice of a single standard would not be obvious. There is again the problem of how to deal with conformationally flexible molecules, although Simon et al [1977] do make some suggestions for dealing with this problem. Clearly, this descriptor is most effectively used with congeneric databases.

STERIMOL. Another method of dealing with steric effects is the STERIMOL approach of Verloop et al [1976]. Their method can be understood by referring to Figure 4. The substituent (in Fig. 4, methyl) is situated in a close-fitting box such that all sides are in contact with the van der Waals envelope of the substituent at at least one point. The two planes forming the ends of the box are positioned to be perpendicular to the bond axis of the substituent bond joining the substituent to its parent structure. Because there may still be several possible rotational orientations of the box about the bond axis, an orientation is chosen that minimizes a perpendicular distance from one side of the box and the bond axis. The dimensions of this box then represent the dimensions of the substituent: L is the length of the box from the point of attachment of the substituent to the end of the box and B1–B4 are the perpendicular distances from each of the remaining four sides of the box to the bond axis, B1 being the minimum and B4 the maximum such distance.

It is clear that one drawback of this method is the number of parameters. In addition, for flexible substituents, it is not always possible to define unambiguously the shape and, therefore, the parameters above. Verloop et al [1976] have tabulated the STERIMOL parameters for a number of substituents in the reference given.

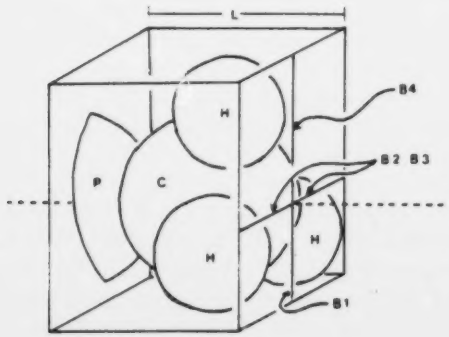


Fig. 4. Verloop steric parameters for a methyl group attached to the parent structure P. For methyl: B1 < B2 = B3 < B4.

Molecular shape analysis (MSA). The molecular shape analysis (MSA) of Hopfinger and Potenzone [1982] is a method that attempts to deal with the variety of conformations a molecule can adopt. It is based on the volume differences between the most stable conformation of the most active compound in a database and the compound being evaluated for activity. The most stable conformation may be determined by molecular mechanics [Hopfinger and Potenzone, 1982; Burkert and Allinger, 1982] or quantum mechanics [Grunwald et al, 1979]. It differs from the previously described MSD in that the molecular volume is actually approximated and in the direct way in which conformations are dealt with. Again, though, it would really be useful only with congeners.

Molar refractivity (MR). Molar refractivity (MR) was introduced along with molecular weights as a first-order estimate of steric effects by Hansch et al [1973]. It can be calculated from some rather basic quantities and the Lorentz-Lorenz equation:

$$MR = \left(\frac{n^2 - 1}{n^2 + 2}\right)\left(\frac{M}{d}\right). \tag{20}$$

where n is the refractive index at the sodium D line, M is the molecular weight, and d is the density. In spite of its crudeness, this descriptor is used fairly often in place of the relatively scarce Taft steric parameters.

Quantum variables. Quantum variables have also been used in QSAR. Theoretically, quantum chemistry can yield any information about reactivity one might want. (Some of the reactivity indices discussed above have been calculated using quantum chemical methods.) In practice, it is usually of more limited use because of computational difficulties, mainly excessive time requirements for the most accurate methods. In that one must make many approximations to bring the calculational time down to reasonable limits, one often sacrifices accuracy. It is still quite a useful method for some problems when used in the proper context, particularly when one is close to understanding specific molecular details of a biological activity. (For an example of a recent use of quantum chemical calculations in sharpening the details of an enzymic mechanism, see Osman and Basch [1984].)

The most common use of quantum chemical methods is more prosaic. A number of quantum chemical variables are included in Hansch-type analyses of biological activity. Some of these variables are atomic charge densities, net atomic charges, net sigma (bond) atomic charges, net π (bond) atomic charges, highest occupied molecular orbital (HOMO) electron densities and lowest unoccupied molecular orbital (LUMO) electron densities, as well as the HOMO and LUMO energies [Osman et al, 1979]. Also included have been the nucleophilic and electrophilic superdelocalizibilities, fancy terms for indices of nucleophilic and electrophilic reactivity [Fukui et al, 1960]. Most of the indices of reactivity can easily be calculated using any of a number of readily available quantum chemical computer programs. (The organization QCPE (Quantum Chemistry Program Exchange) at Indiana University provides a number of these programs at cost.) There are many levels of sophistication in these calculations. For the purposes described here, the level of sophistication most practical is probably at the semiempirical level consisting of MINDO/3, CNDO/2, MNDO, and perhaps the Huckel and extended Huckel. These programs have been written for a variety of computers, including, in some cases, personal computers.

Limitations

Colinearity. Hansch [1976] and others [Craig, 1971; Lewi, 1980] have pointed out that a degree of colinearity may exist between the various parameters of the linear-free energy relationships discussed. This creates a number of problems. Most importantly, depending on the degree of correlation, it may subvert the underlying assumptions of the particular statistics used for the analysis. It clearly would also cause problems in interpretation. Lewi [1980] has pointed out that it may cause unstable solutions in linear regression analysis (the most often used of the statistical analyses), ie, that the results will not be reproducible from run to run. Craig [1971] has actually demonstrated colinearity for a number of the Hansch parameters. He found that the substituent parameter, π, measuring hydrophobicity, is basically independent of the so-called "polar" or electronic terms F, R, σ, and E_r. The term π was clearly correlated to molecular volume and somewhat less strongly correlated to E_s (steric parameters) and MR (group refractivities). The fact, however, is that colinearity cannot, in general, be predicted. Hansch [1976] and Craig [1971] recommend the plotting of pairs of variables to be used in an analysis (the "Craig plots") to determine the degree of colinearity. There are other, more sophisticated methods for dealing with this problem, such as the use of data analysis techniques before statistical treatment of the data [Hansch, 1976; Lewi, 1980].

Congenericity requirement. Another more fundamental problem in using the Hansch method as applied to the problem of mutagenicity (and carcinogenicity) is the fact that this method basically requires congeneric databases. Also, the sheer numbers of compounds tested for mutagenicity and carcinogenicity do not lend themselves to this type of analysis [Hansch, 1978]. Given the diverse nature and mechanisms of action of mutagens, genotoxicants, and carcinogens, the extrathermodynamic methods will find their greatest usefulness only as we come to understand more of the details of these molecular mechanisms or in studying relatively small, congeneric databases.

SIMCA Method

The SIMCA (one meaning of the acronym is "soft, independent modeling of class analogy") method of Wold and Sjostrom [1977] is one of a variety of methods employing Hansch-type variables and the pattern recognition method of mathematical analysis. As in all of the methods described, one part of the analysis consists of defining the compounds of a training set as points in a multidimensional space. The vectors representing the points consists of variables thought to be related to the activity of the compounds, such as the Hansch physicochemical variables above. The mathematical treatment that the SIMCA method uses is the principal component method. The concept of principal component methods is illustrated in Figure 5. In this simple case, we have two classes of "compounds," with different biological activities. On each axis of the graph, we have one kind of physicochemical property. As in any successful resolution by this method, the two classes of compounds have ended up rather well separated as to class. Within each class, an equation can be developed that fits the activity using "components," $u_{ak}{}^q$, of the property variables:

$$y_{ik} = m_i{}^q + \sum^{A} b_{ia}{}^q \cdot u_{ak}{}^q + e_{ik}{}^q. \qquad (21)$$

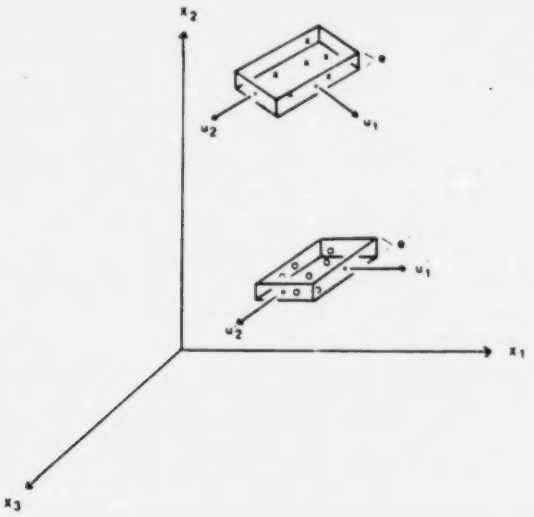


Fig. 5. Illustration of the principal component method used in SIMCA. Each axis represents a different physicochemical property. The ideal situation is shown in which the two sets of compounds (x and o) are well separated and their biological activities can be represented by equation 21. The u's represent the principal components and the e's the error quantities. A = 2 for each case.

where y_{ik} is the property variable for property i and object k in class q, m_i^q is the mean of the variable i in class q and with b_{ia}^q defines the position and direction of the class, u_{ak}^q defines the position and direction of the particular object k within the class (in terms of a composite of the property variables), and the e_k^q are the error quantities. The quantity A is the dimensionality needed to locate the objects in the property space (a dimension of two is indicated in Fig. 5). This fact, that one obtains an equation relating activity to the parameters that have been input, makes this method somewhat different from other pattern recognition methods, which are basically classification schemes. SIMCA also has a relation to factor analysis, which uses an equation similar to 21 but which does no preliminary classification, working only on the training set as a whole. To summarize, SIMCA classifies the various similar compounds based on activity class first and then derives an equation for each class describing the data structure within that class.

Other Methods Using Hansch Variables

Methods using Hansch variables and cluster analysis statistical method. A number of authors have used one or more variants of cluster analysis in QSAR. Cluster analysis is more of a classification scheme than a method of allowing actual predictions of activities. Its use is really in identifying the factors important in classifying a compound's activity or lack of it. Hansch and Unger [1973] have also used cluster analysis to identify redundancies among substituent parameters for a given set of substituents. First, a space is defined, in the case of the Hansch-Unger

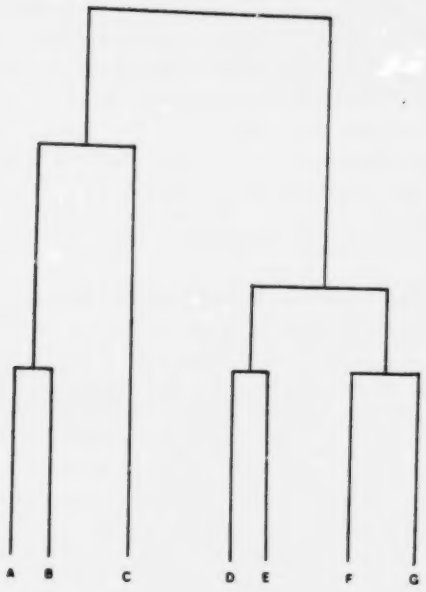


Fig. 6. Graphical representation (tree diagram) of cluster analysis method.

report [1973] this space is the space of K kinds of substituent parameters (σs, πs, etc). Then, each of the N substituents is defined in that space by each of its K substituent parameter values. In the hierarchal clustering method, the analysis begins with the evaluation of the Euclidean distance, d_{ij}, between each pair of points i and j in the space. The closest two points are clustered into a new pseudopoint, and the distances between pairs are calculated again:

$$d_{ij} = [\sum^{N} (x_{ik}' - x_{jk}')^2]^{1/2} \quad i,j = N, \tag{22}$$

where the x_{ik}'s are scale-corrected substituent parameter values for the ith point and the kth parameter and the x_{jk}'s are the cluster averages (or single points on the first run) for the jth point and kth parameter type. The scaling is done through the following equation:

$$x_{ik}' = (x_{ik} - x_i')/s_i, \tag{23}$$

where s_i is the standard deviation of the sets of k parameters. The process is repeated until one large cluster is formed. This is not necessarily the most useful level of clustering. The most useful level of clustering is clearly that in which the points in each cluster are most similar and the differences between clusters are greatest. (The tree diagram in Fig. 6 represents the results of the process pictorially.) This level may

be established by a methodology that allows one to evaluate the information loss or gain as a function of the number of clusters [Hansch and Unger, 1973; Tyron and Bailey, 1970]. The method has been used in other contexts also. One interesting use focused on clustering indicator variables (indicating the presence or absence of a particular structural fragment, vide infra) with certain physicochemical properties, ie, associating a particular substituent or structural fragment with the physical property that contributes the most to the overall biological activity [Dunn et al, 1976]. The method also lends itself to the more common use of pattern recognition methods in this area by looking for the physicochemical properties associated with a particular biological activity.

Hansch variables and discriminant analysis. Linear discriminant analysis (and the linear learning machine) are also pattern recognition methods that have recently found use in QSAR work. These two methods use vector dot product multiplication to maximize distances in the variable space between compounds that belong to different classes. The variables may again be physicochemical properties, but indicator variables have been used [Chu et al, 1975]. The vector equation below is optimized at each point in the space to give the largest possible positive and negative C_{ik}, ie, in an absolute sense. Optimizing in these two directions gives a "decision surface," ie, boundary surface between the two classes of compounds. The vector equation is:

$$S = \sum \vec{W}_i \cdot \vec{C}_{ik} = \sum W_i \cdot C_{ik} \cos\theta_{w,c}, \tag{24}$$

where W is the "weight" vector and is perpendicular to the decision surface and defines it, C_{ik} is the vector representing the various compounds k in the parameter space of dimension N, and $\theta_{w,c}$ is the angle between the vectors W_i and C_{ik}. The end result can be visualized easily in two dimensions as in Figure 7. At this dimensionality, the decision surface is defined by a line, and the points representing the various compounds are scattered to each side of it. A new compound's activity class is defined by which side of the decision surface it lies on.

When this particular method is combined with the mechanics that allow updating in the light of new data, it is referred to as a *learning machine*. When a new compound is tested and found to be incorrectly placed on one or another side of the decision surface, the above equation is corrected until the new compound is correctly placed.

There are other methods for generating decision surfaces, although the above is simplest. Martin et al [1974] used discriminant analysis and another method of decision surface manufacture to define the relationship between physicochemical properties and monoamine oxidase activities.

Some of the statistical expertise required to do these kind of analyses is available from the University of California, Los Angeles, through their BMD-UCLA Biomedical Program BMDP7M [Dixon, 1977] or the Statistical Analysis System (SAS) package from North Carolina State University in Raleigh [Service, 1972].

Although publications analyzing mutagenic or carcinogenic activity are scarce at present, it appears that discriminant analysis would be especially useful in the analysis of mutagenic and carcinogenic activity; much of the published work is not quantitative but merely indicates activity or nonactivity, eg, the reports of the GeneTox Program [Waters, 1979; Waters and Auletta, 1981; Palajda and Rosenkranz, 1985].

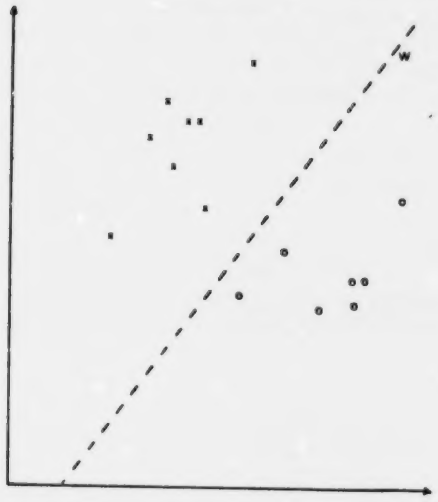


Fig. 7. Two sets of compounds (x and o) of differing biological activities separated by the decision surface (line W) of equation 24.

CONNECTIVITY METHODS

In this section, a number of methods are discussed that treat biological activity and molecular structure from a perspective different from that discussed above. These methods are actually complementary to those methods that relate biological activity to physical properties. In many cases of interest (especially among some of the new compounds whose potentials for mutagenicity or carcinogenicity are most important), the physical properties of the compounds and their substituents are unknown or else the data representing the biological activity are not quantitative. Most of the strategies that have arisen to get around this problem do have certain features if not philosophies in common. The most important of these is that somehow the actual molecular architecture is ultimately responsible for the observed biological (and chemical) properties. The differences between these connectivity methods reduce to different representations of the structures and the subsequent statistical treatment of the data.

Free-Wilson Method

The Free-Wilson method grew out of a direct response to the lack of physical and biological data on compounds whose biological activity was of interest [Free and Wilson, 1964]. The authors were particularly concerned about the gaps in biological testing data that would help establish quantitative structure-activity relationships. It was assumed in the original development of their method that the effect of a substituent on biological activity (BA) was additive and independent of the presence or absence of other substituents at other positions. The generalized form of the equation they proposed is then:

$$BA = \Sigma A_{ij} \cdot S_{ij} + k, \tag{25}$$

where A_{ij} is the activity contribution of substituent i at site j and S_{ij} is an indicator variable given the value 1 if the substituent i is present at site j and 0 if it is not. The constant k is the average value of the biological activity and is assigned to the hypothetical structure that has neither substituents nor hydrogen in the sites of interest. The equation can then be solved in a least-squares manner. The consideration of a concrete example might help make clear the assumptions and use of the model.

Consider the substituted naphthalene (4 in Fig. 1). We assume it has three possible sites of substitution, X, Y, and Z. In addition, site X has three variations, Y has two, and Z has two for a total number of variations of 12. There are a total of eight unknowns, but the assignment of the average value of the biological activity to the unsubstituted structure means that we have the three additional equations assigning an overall average contribution of 0 to each site by all substituents:

$$0 = \Sigma A_{ij}. \tag{26}$$

This means that we need a minimum of five compounds with their measured biological activities to solve for the A_{ij}. Because of variations in measured biological activities, it is usually better to have more than the minimum required data. In Table II the hypothetical data are summarized.

It is interesting to note that an equivalence can be demonstrated between the linear-free energy relationships of Hansch and the Free-Wilson method [Osman et al, 1979]. The Free-Wilson method clearly is restricted to congeneric databases, and this fact limits its usefulness in the kinds of problems considered by environmental scientists. The method is important, however, in demonstrating that the additivity of substituent contributions might be a valid assumption in some situations. Many of the more useful models that follow do incorporate in some way this additivity hypothesis.

ADAPT (Automated Data Analysis Using Pattern Recognition Techniques)

Description including pattern recognition statistics. Stuper et al [1977] developed one of the first "no-model" approaches to SAR that is based on the similarity of a test compound to members of a training set. Similarity is defined by a number of descriptors, which will be discussed below. It is also one of the first methods that can be applied to relatively large, noncongeneric databases. The method has been dubbed *ADAPT* (for automated data analysis by pattern recognition tech-

TABLE II. Hypothetical Data Illustrating Free-Wilson Method

	Variation							
	Site X			Site Y		Site Z		Biological
Compound	H	OCH_3	NO_2	H	Cl	H	Br	activity
I	*				*		*	A_I
II		*		*		*		A_{II}
III			*	*		*		A_{III}
IV		*			*		*	A_{IV}
V			*		*		*	A_V

*Presence of substituent at that site.

niques). The program was originally designed to run on the MODCOMP/II/25 16-bit computer. There are several published reviews of the method in varying detail [Stuper et al. 1977, 1979; Jurs et al. 1978].

Descriptors. As in the previously discussed SIMCA pattern recognition method, each compound of interest is described as a point in a multidimensional variable space. In this case, the variables are not always physicochemical parameters but such varied items as fragment descriptors, substructure descriptors, molecular connectivity descriptors, environment descriptors, path and path environment descriptors, and geometrical descriptors. The atomic charges, partition coefficients, and molar refraction can also be included.

Fragment descriptors. These descriptors include atom type counts, bond type counts, the molecular weights of the compounds, the number of ring atoms, and the number of rings.

Substructure descriptors. This descriptor is supplied by the researcher at the time of the analysis. It may represent a molecular connectivity descriptor. The molecular connectivity is an index of the relative branching of a molecule and depends on the hydrogen-suppressed formula or graph of the molecule. For instance, although it may be clear that 2,2-dimethylbutane is more branched than n-hexane, how is 1,1-dimethylcylopentane to be compared to 1,2-dimethylcyclopentane? Molecular connectivity indices are one attempt to answer such questions in an objective fashion. These indices have been shown to be respectable correlatives of boiling points [Randic, 1975]. There are a number of different orders of molecular connectivity: path 1 ($^1\chi$) connectivities are obtained by adding bond contributions (δ_i) whose size is determined by the number of nonhydrogen ligands attached to each atom i and j of the bond:

$$^1\chi = \Sigma(\delta_i \cdot \delta_j)^{-1/2}. \tag{27}$$

In the case of propane, for example, the δs are equal to 1 for carbon 1, 2 for carbon 2, and 1 for carbon 3. This gives two path 1 contributions of 0.7071 for a $^1\chi$ of 1.4142 [Randic, 1975]. The higher-order contributions (path 2, etc) are obtained by considering bonds further than just one bond length [Murray, 1977], for path 2:

$$^2\chi = \Sigma(\delta_i \cdot \delta_j \cdot \delta_k)^{-1/2}, \tag{28}$$

and, in general, for linear path lengths p:

$$^m\chi_p = \sum_{s=1}^{nm} \prod_{i}^{m+1} (\delta_i)_s^{-1/2}. \tag{29}$$

where n_m is the number of paths with m edges and s identifies a particular subgraph. Extensions to heroatoms have been devised [Hall and Kier, 1977; Kier and Hall, 1976a,b].

Environment descriptor. This descriptor is generated from the entered substructure descriptors. The environment of a substructure is generated by doing a path 1 molecular connectivity calculation over the substructure and its nearest neighbor atoms in a molecule in which it is found embedded.

Partial atomic charges. The atomic charges are generated by the Del Re [1948] method, which is formally similar to the Huckel method but deals only with the sigma framework of the molecule. Searches are conducted via average atomic charges over an entire substructure or a specific atom within a substructure.

Partition coefficients. These parameters are estimated by the Hansch method [Leo et al, 1975] and have been discussed in connection with the extrathermodynamic methods.

Geometric descriptors. These resemble the steric parameters of the extrathermodynamic approach (except that they are calculated for the whole molecule rather than just for a substituent) and include molecular volume, principal axes, and ratios of principal axes of the molecules.

Statistical analysis routines. Once the descriptor generation steps are finished, the user of ADAPT must choose those descriptors of the rather extensive list above that will be submitted for statistical analysis. (In one example cited, Jurs et al [1978] pointed out that over 50 parameters were developed for each of 200 compounds in a database analyzed.) An equation defining the compound as a point in the chosen parameter space can be written:

$$X_a = x_1 i + x_2 j \ldots + x_n k, \tag{30}$$

where X_a is the compound a's vector in the parameter space repesented by the nx components.

The ADAPT program has two kinds of statistical packages, described by Jurs et al [1978] as parametric and nonparametric. The parametric routines are based on mean vectors, covariance matrices, or other statistical measures of the distributions (classes) being separated. The nonparametric routines are based on other measures of similarity, such as the Euclidean distance method described above in connection with cluster analysis. Most of the various methods used here are to obtain an optimal discriminant function (ie, decision surface, the more colorful term used in the discussion of linear discriminant analysis above) between two classes of objects (in our case, compounds). In conjunction with these programs is an updating routine that allows the discriminant function to be updated in light of new data. Thus the program is heuristic. This is an especially important feature in the environmental field. Those interested in the details of the procedures will want to consult the reference works used in the development of the ADAPT program [Tou and Gonzalez, 1975; Jurs et al, 1969; Morgan and Deming, 1974; Ritter et al, 1975]. In addition, the actual statistical and pattern recognition routines used were obtained from Dixon [1977].

The last step in the ADAPT analysis of a database is feature selection. This is done interactively through several routines using variance selection techniques [Zander et al, 1975].

The experience of the Jurs group in using ADAPT and its various statistical routines has led them to conclude that for descriptors that are not continuous variables (fragment descriptors, etc.) or for compound (point) distributions that are not normal in the statistical sense, the nonparametric methods are better than the parametric ones. Further, they state that the methods making use of the decision surface are superior to Euclidean distance methods for these cases also [Jurs et al, 1983].

CASE (Computer Automated Structure Evaluation)

General description. One problem that remains unaddressed is the fact that most of the paradigms described so far require a great deal of input from the researcher and, hence (from one perspective), a great deal of bias. This is particularly true in terms of the choice of descriptors. It would be useful to have a method that made only the most basic assumptions about the relationship between molecular structure and biological function. The computer automated structure evaluation (CASE) [Klopman, 1984] is one attempt at doing this. This program was designed to be as automatic as possible and to introduce as little bias as possible. It presently operates on a Digital Equipment Corporation VAX 11/750 computer.

The descriptors. The descriptors are based on the substructures (fragments) of the molecules of the training set. The fragments are generated by decomposing each molecule of the database into all possible pieces consisting of from three to ten contiguously connected heavy atoms together with their attached hydrogens. (Clearly, this generates a number of redundancies; these are dealt with at a later stage of the analysis.) Each fragment also possesses three labels, one defining bond multiplicities, the other indicating the presence of a functionality, such as $-NH_2$, $-COOH$, etc, and the last specifying whether the fragment originated from an active or inactive molecule.

The analysis. Before any further analysis, the redundant fragments are removed. A statistical analysis of the remaining fragment distribution in each of the two pools, inactive and active, is made. Assuming a binomial distribution, any fragment is considered irrelevant to the biological activity of interest if its distribution among the actives and inactives is similar to its distribution among the total sample of molecules. A fragment is considered relevant to the biological activity if its occurence in the distribution is $<5\%$ probable on the basis of a chance (binomial) distribution. Such a fragment retains the activity label of its origin. A concrete example should make clear the process.

Consider that the activity of interest is acidity as determined by the property of turning blue litmus paper red. A sample of 100 compounds was tested for this demonstration; 40 of them were acids and the others were not by this test. The compounds were coded, entered, and submitted for analysis. The resulting fragment analysis is presented in Table III. A fragment like CH_2-CH_2-CH_2- is found in 63 molecules, 28 acids, and 35 nonacids. This is a distribution one would expect for a random sample. On the other hand, the fragment O=C-OH is found in 38 acids and four nonacids, a distribution not at all likely to have occurred by chance. The presumption is then that O=C-OH is related to the property of acidity as defined. The discrepancies observed (four molecules possessed the grouping but did not express acidity) may be the result of inaccuracies in measurement or the presence of inactivating fragments possessing $-NH_2$ (ie, a zwitterion). Although this result might seem trivial, it has become evident that, where biological activity is concerned, it cannot be reduced to the simple 1-4 heavy atom sequences of organic chemists. In fact, our analyses of real problems suggest that biological "functionalities" consist of considerably larger groupings of atoms. An additional feature of the CASE program, difficult to take into account by mere inspection, is synergisms between fragments.

TABLE III. CASE Analysis of the Blue Litmus Database: Sample Fragments for Analysis of Acidity of Organic Acids*

Fragment	Total No.	No. present in inactives	No. present in actives
CH_3-CH_2-CH_2-	25	16	9
CH_2-CH_2-CH_2-	63	35	28
CH=CH-CH_2-	11	7	4
C''-CH_2-CH-	14	7	7
O=C-CH_2-	55	25	30+
O=C-OH	42	4	38+++
CH=C-OH	8	2	6+
CH_2-CH-Cl	10	6	4
NH_2-CH-CH_2-	7	6	1−
NH_2-CH-C=	3	3	0−

*Total number of molecules, 100; number of actives, 40. From Klopman [1984], with permission.

The program does a pair-by-pair analysis of fragments to determine if any fragment has a greater potency in conjunction with another fragment than it does alone.

Testing new molecules. After the training of the program for a particular database is complete, one would then like to test new compounds of unknown activity. Submitting such a structure to analysis (with a key indicating that the activity is unknown) results in the fragments of it being compared to the fragments of relevance obtained from the training set. If overlaps are found, a probability that the new compound possesses the activity may be calculated based on the probability that the overlapping fragments were relevant to activity in the training set. If the actual activity of this (or another compound) becomes available, then the training analysis is performed again with the new data, thus updating and correcting the program's projection ability. It thus has the characteristics of a learning machine.

Quantitative aspects. The program also has the ability to estimate potencies based on fragments, logPs and $(logP)^2$s. Additivity of fragment contributions to the activity is assumed:

$$A_m = \Sigma a_i f_i + \log P_m + (\log P_m)^2 + k, \qquad (31)$$

where A_m is the activity calculated for molecule m, a_i is the activity contribution of fragment i, and f_i is either an indicator variable or the number of occurrences of fragment i in molecule m (the choice is a program option). A linear multiple regression analysis is performed with the above equation using the partial and full F tests to determine whether a variable should be included [Bevington, 1969]. Recently, criticisms of the multiple linear regression analysis have pointed out that many of the standard tests of significance do not properly reflect the probability of obtaining chance correlations (especially when "large" numbers of variables are screened) [Topliss and Edwards, 1979; Wold and Dunn, 1983]. The CASE method was especially sensitive to this criticism; it creates and screens independent variables (fragments) numbering from 50% to 100% of the total number of dependent variables (compounds' activities).

The general problem of causality is one of great concern to those of us involved in QSAR methods and was investigated experimentally by us recently, especially but

not exclusively as it relates to the CASE program [Klopman and Kalos, 1985]. It was found that as a general rule the safeguards that are built into the CASE program, ie, eliminating redundant fragments (the colinearity problem, which is one facet of the problem), the accepting into the regression analysis of only those fragments of a high probability of being associated with activity, and the inclusion of partial and full F tests [Bevington, 1969; Weissberg, 1980], are generally sufficient to prevent chance correlations.

The Enslein Approach

The Enslein group has developed an approach similar to the CASE program in its use of substructure fragments [Enslein and Craig, 1978, 1982; Enslein et al, 1983a,b]. However, the Enslein approach differs from the CASE methodology by possessing a "dictionary" of fragments. The CASE program has an open-ended set of substructure fragments. By *open-ended*, we mean that there are no restrictions placed on the choice of fragments representing the activity/inactivity of a compound. It might be *any* fragment the molecule possesses as long as it possesses at least three and no more than ten heavy atoms. In the Enslein procedure, the activating and inactivating fragments must come from the dictionary of available fragments, and although this is an extensive list, it could be that just the substructure one needs to predict activity or inactivity is not there. The Enslein procedure also differs from CASE in the set of statistical routines to evaluate the importance of any given substructural key and in the representation of those keys. In addition, with Enslein's method, some keys may be marked or prioritized to encourage the statistical packages to include certain substructures.

The dictionary of substructural keys or fragments is obtained from the Wisswesser line notation as found in the CROSSBOW program [Eakin et al, 1974] with some modifications. The original CROSSBOW set of keys consisted of only 148 substructures; Enslein has added others to improve representation and to include certain types of fragments suspected of being related to specific biological activities, such as carcinogenesis [Enslein et al, 1983b]. The total number of substructural keys at last count was 350.

The model building consists of, as a first step, the cluster analysis procedure of Hall and Khanna [1977]. This separates the database into structured clusters or classes and may indicate outliers. The next step is the prioritization of the substructural keys and other descriptors (which can include physical properties and the molecular weight), which compels the discriminant analysis routine to accept higher priority features first, such as those suspected of being more relevant to activity, and apparently results in a better-quality discriminant equation. Next begins the stepwise discriminant analysis procedure, which is the heart of the program, identifying which of the substructure or other descriptors are to be considered important for activity or lack of activity. This set of descriptors, coming out of the discriminant analysis, is then scrutinized, and those descriptors contributing least to the separation of active from inactive compounds are removed one by one until the remaining descriptors meet a predetermined inclusion criteria. The next step is a ridge regression procedure [Marquardt and Snee, 1975], which eliminates descriptors that might cause instabilities in the discriminant equation. As a last step, all the surviving descriptors are submitted to another discriminant analysis. This results in a final set of descriptors and their associated coefficients.

There are two sets of coefficients, one for activating features and one associated with inactivating features. One obtains a projected probability, P, of activity for a given compound by combining cumulative discriminant scores for the activating features and the discriminant scores for the inactivating features in the following equation:

$$P = \frac{\exp(\text{activating score})}{\exp(\text{inactivating score}) + \exp(\text{activating score})} \quad (32)$$

Confidence levels can be assigned to this estimate based on whether, for instance, only a few features of the compound are found to be represented in the original database (training set).

The accuracy of the discriminant analysis can be evaluated in a number of ways. The most formalized of the procedures is the two-sample Kolmogorov-Smirnov test [Seigel, 1956]. This test is based on the idea that the distribution of the misclassified compounds must conform to a binomial distribution centered on the probability of 0.50. In addition, test compounds can be submitted for prediction for which the true activity is known and which were not included in the training set.

Other Related Methods

There are a number of methods related to those described above. Another methodology closely related to the CASE and the Enslein methodology is the paradigm developed by Hodes et al [1977]. It represents molecules as formed from augmented atoms, which are structural fragments consisting of a central atom and all its attached atoms. Then, the augmented atoms of importance are identified by the extent their distribution deviates from a binomial distribution. The formulas used by the methods for establishing a probability of associating a fragment with activity are different from CASE, as is the fact that the CASE method can be used to predict a potency as well as a probability of activity. An extension of the Hodes methodology was developed by Tinker [1981], which permitted use of semiquantitative activity categories and defined whether the augmented atoms were part of a ring or not. In its predictive mode, this method was able to place correctly (within one activity category) 88% of 34 compounds of potential carcinogenic activity.

A method utilizing augmented atoms, heteropath fragments, and ring fragments in conjunction with pattern recognition methods (learning machine and nearest neighbor methods) was used in the analysis of 138 agents tested for antitumorigenicity [Chu et al, 1975]. The heteropath fragment is defined as the molecular subunit consisting of two heteroatoms and the intervening chain of carbon atoms ; the ring fragments involved definition of ring size and presence or absence of heteroatoms and their placement and substituent patterns for six-membered rings. The methods were able to classify correctly the 138 compounds, and each pattern recognition method did fairly well in classifying test compounds not in the training set (although the test set in this case was relatively congeneric).

SAR AS APPLIED TO THE STUDY OF MUTAGENICITY AND CARCINOGENICITY

Early Studies on Structure-Activity Relationships of Carcinogens

Although present-day structure-activity studies are concerned primarily with the relationship between structure and mutagenicity (or genotoxicity), this is a rather

recent development that resulted from the rationalization that many carcinogens derived their activity from an ability to react with and modify DNA (ie, genotoxic carcinogens) and from the relative ease with which mutagenicity and genotoxicity data could be generated. However, prior to this understanding, most SAR studies, although they were more simplistic, were concerned with animal carcinogenicity investigations. One of the earliest modern studies dealt with the role of the methylene in the carcinogenicity of 2-acetylaminofluorene (5a in Fig. 1): Miller et al [1949] replaced the $-CH_2-$ bridge with an -S- (ie, 3-acetylaminodibenzothiophene: 5b), S=O (ie, 3-acetylaminodibenzothiophene-5-oxide; 5c, and -O- (ie, 3-acetylaminodibenzofuran; 5d) and finally deleting the $-CH_2-$ bridge altogether (ie, 4-aminobiphenyl). It was thus found that the -S- -substituted analog did not differ from 2-acetylaminofluorene in its ability to induce mammary and ear duct cancers, whereas the analog containing S=O or -O- exhibited decreased activity. Moreover, none of the three analogs retained the ability to induce hepatic tumors. 4-Dimethylaminobiphenyl, on the other hand, induced all three kinds of tumors.

Contemporary with these were studies on 4-aminostilbene (6 in Fig. 1) and its congeners [Haddow et al, 1948]. These studies suggested the requirement of planarity for carcinogenicity as evidenced by the requirement of an unbroken conjugation of the amino group with both ring structures leading to a quinoid configuration. Miller et al [1949] used this as the basis for explaining the greater carcinogenicity of fluorene than of the biphenyl derivatives. The presence of the methylene bridge induced a significant shift in the ultraviolet spectrum towards the longer wavelengths, which was explained as reflecting the fact that the bridge increased charge transfer between the two rings as a consequence of the greater planarity of the ring system [Sandin et al, 1952]. This led to the prediction that those carcinogens that absorb maximally at longer wavelengths would be more carcinogenic. This was indeed found to be the case among the fluorene and biphenyl derivatives studied [Sandin et al, 1952].

Early studies on SARs of carcinogens were hampered by the fact that the relationships between structure, biotransformation, and carcinogenicity were not well understood. Still, the work of the earlier investigators deserves description; they have laid the foundation for understanding the SAR basis of carcinogenicity. Thus, Bergmann [1942] realized that carcinogenic polycyclic aromatic hydrocarbons (PAHs) shared a flat, planar configuration, which when lost was accompanied by a decreased carcinogenicity. The decrease in planarity could be due to hydrogenation [Shear and Leiter, 1941], alkylation (Badger, 1948), or a shift from the *trans* to the *cis* configuration as for aminostilbene [Haddow and Kon, 1947].

Because the ease of formation of picrate is related to planarity, the stability (melting point) of picrates formed by PAHs was used as a predictive measure of carcinogenicity [Orchin, 1951]; this did not, however, hold for all PAHs [Takemura et al, 1953].

Robinson [1946] observed that carcinogenic PAHs shared a phenanthrene type of bridge that could be activated by substitution or by the addition of extra benzene rings. Further, benzene possesses a hexagonal structure, with each carbon having a valence of four and each hydrogen a valence of one. There are thus 30 valence electrons. Twelve of these are involved in carbon-to-carbon bonds and 12 in the six carbon-hydrogen bonds, leaving six "unassigned," the pi-electrons. The disposition of these can be accounted for by quantum mechanical considerations [Coulson, 1953]. Thus, free valency numbers can be assigned to the various carbon atoms of cyclic structures. These, in turn, are measures of the unused bonding ability of the carbons. Whereas in benzene, each carbon, being identical, has the same valency number, in

PAHs this is not so. There are bond orders and relative double-bond characters. For example, in benz[a]anthracene (7 in Fig. 1), the carbon 7 has the greatest free valence number and the 5,6-bond the greatest bond order [Badger, 1954]. Those positions with the greatest free valence number are more readily attacked in substitution reactions. Thus, the presence of the phenanthrene-type bond of high order may be a requisite for carcinogenicity. The most readily substituted positions are those with the greatest free valence number, as was shown to be true for the methylated benz[a]anthracene isomeric series [Badger, 1954], ie, the 7, 8, and 12 positions, having the greatest conjugation activity and methylation at these sites yielding the most carcinogenic species [Badger, 1954; Newman and Kosak, 1949]. This, in turn, led to a correlation between the site of methylation, carcinogenic potency and ultraviolet absorption (ie, magnitude of the bathochromic shift), and conjugating ability of the site at which alkylation occurs that can be used in a predictive fashion [Badger et al, 1952; Jones, 1940].

These realizations led to the concept of a correlation between reactive positions and carcinogenicity resulting from conjugation at these positions. Thus methyl groups, which are electron donors, contribute a higher electron charge to the PAH at reactive positions than at positions having lower conjugating capacity, thereby inducing greater carcinogenicity. Accordingly, a relationship can be postulated between carcinogenicity and the electronic configuration as well as the electron density.

These findings led the Pullmans [A. Pullman, 1947, 1948; B. Pullman, 1946, 1948; Pullman and Pullman 1946, 1948] and the Daudels [R. Daudel, 1948; P. Daudel, 1950; Daudel and Daudel, 1950] to calculate that benz[a]anthracene and benzo[c]phenanthrene as well as chrysene had phenanthrene-type bonds with particularly high electron densities, whereas noncarcinogens such as triphenylene or napthacene do not possess such regions. The electron density at phenanthrene bonds would be expected to increase methylation. The phenanthrene type of bond was called the *K region* (Fig. 8) for "Krebs" (ie, cancer) [Schmidt, 1939a,b]. It was suggested that the K region was associated with carcinogenic potential, an optimal pi-electron density above a critical value being required for carcinogenic activity [Badger, 1954; Greenwood, 1951]. Perbenzoic acid was postulated as being an agent capable of reacting with the K region PAHs at a rate proportional to the carcinogenicity; hence this rate was postulated to be predictive of ultimate carcinogenicity [Boyland, 1950]. Another reagent specific for the double bond K region PAHs and therefore presumably predictive of carcinogenicity is osmium tetroxide [Badger, 1949, 1950; Badger and Lynn, 1950; Cook and Schowenthal, 1948].

These findings led to the suggestion that the carcinogenicity of PAHs was dependent on planar, reactive double bonds and methylation at positions favorable for

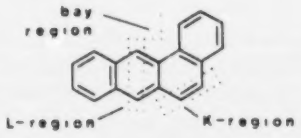


Fig. 8. The bay, L, and K regions of polycyclic aromatic hydrocarbons (PAHs).

further activation of the double bonds. The Pullmans [1955] suggested further that the carcinogenicity of tetra- and pentacyclic PAHs was related to the presence of another region of low activity (the L region; Fig. 8). For carcinogenicity, some of the free valences of the two carbons consisting of the L region had to be less than one. Finally, it is only thanks to the work of the Millers, which identified the electrophilic nature of the ultimate carcinogen, and the work of Jerina and associates on the requirement of the bay region that the structural features responsible for the carcinogenicity of PAHs (ie, formation of diol epoxides) was recognized. The potential for forming bay region epoxides capable of forming carbonium ions can be predicted from quantum mechanical considerations [Jerina and Lehr, 1977].

Relevant Studies Explicitly Using the Hansch Method

Studies relating to the rate of metabolism of compounds by microsomal preparations. A number of Hansch-type structure-activity studies have been applied to rates of metabolism by various microsomal preparations. Because the *Salmonella* mutagenicity test involves use of one or more microsomal preparations, it seemed the results of such analyses might be of some interest. The study of Martin and Hansch [1971] looked at rat liver microsomal NADPH oxidase acting on drugs containing the amino functionality. A correlation equation was developed between logP and the log of the Michaelis constant, K_m. The correlation was poorer for ionized species, and no optimal lipophilicity was detected. Yih and van Rossum [1977a] attempted to correlate the rate of microsomal oxidation to dissociation constants and $(\log P)^2$ for the barbiturates (8 in Fig. 1). The finding here was that a parabolic relationship suggesting an optimal lipophilic character held up to the point that the alkyl side chains R began to effect the reaction rate through steric effects. Gammans et al [1977] studied the microsomal hydroxylation of the *trans*-4′-halo-4-acetamide stilbenes (9 in Fig. 1) as a function of logP and found wide variation among activities as measured by K_m, with only small changes in lipophilicity in the halogen series. Yih and van Rossum [1977b] looked again at barbiturates, but this time in a perfused rat liver, and attempted to correlate logP to the rate of clearance of the barbiturates from the blood stream. The correlation was again good except for the halogenated derivatives, which suggests a special metabolic pathway for them. The general conclusion to come out of such work then is that microsomal metabolism is sensitive to the hydrophobic nature of the substrates (with the exceptions noted).

Studies relating to mutagenicity and carcinogenicity. Some of the studies using the Hansch method that apply directly to the question of mutagenicity are briefly discussed here. Hansch's first entry in this area actually appeared with the first paper to describe thoroughly his QSAR method in 1964. This paper described the applications of this method to a number of biological activities including the carcinogenicity of the aromatic amines represented by 10 in Figure 1. The following regression equations were derived by Hansch's treatment for each of the rings, a and b, respectively:

$$\log A_b = -2.02\pi^2 + 0.93\pi + 0.81 \qquad (33)$$
$$r = 0.699$$
$$s = 0.241$$

$$\log B_a = -2.06\pi^2 + 0.94\pi - 0.12\sigma + 0.83 \quad (34)$$
$$r = 0.710$$
$$s = 0.210.$$

What these equations were interpreted to mean was that substituents on the second ring had little *electronic* effect on carcinogenicity, since σ was not picked up by the regression analysis. The main effect of the substituents on the second ring appeared to be in modifying the lipophilicity, as is shown by the more substantial coefficients of the π and π^2 terms in equation 33.

Hansch et al [1979] looked at the SAR between 1-(X-phenyl)-3,3-dialkyltriazenes (11 in Fig. 1) and mutagenicity. The same class of compounds had been looked at earlier as antitumor agents. Two regression equations were derived for the mutagenicity and antitumorigenicity data. Equation 35 represents the result for mutagenicity:

$$\log (1/C) = 1.09\log P - 1.63\sigma^+ + 5.58 \quad (35)$$
$$n = 17$$
$$r = 0.974$$
$$s = 0.315.$$

The quantity C in equation 35 is the molar concentration of triazene giving 30 mutations/100 million bacteria (strain *S typhimurium* TA98). In that this class of compounds requires activation to be mutagenic, the importance of the logP term, as indicated by its substantial coefficient of 1.09, is not surprising. Lipophilicity is often an important property in microsomal oxidation, as was discussed above. The importance of the σ^+ (electron donation) might be a reflection of the occurence of a first step like the following one (Scheme I) in a proposed bioactivation process for conversion of the triazene to a penultimate carcinogen, eg. R^+ [Hansch et al. 1979].

$$XC_6H_4\text{-}N{=}NN\langle^{Me}_{R} \xrightarrow{[o]} XC_6H_4\text{-}N{=}NN\langle^{CH_2OH}_{R} \rightarrow$$
$$XC_6H_4\text{-}N{=}NN\langle^{H}_{R} + CH_2{=}O \rightarrow XC_6H_4\text{-}NHN{=}NR \xrightarrow{H_2O}$$
$$XC_6H_4\text{-}NH_2 + [RN{=}NOH] \rightarrow R^+ \xrightarrow{Nuc}$$

Scheme I.

The following equation was derived for the antitumor properties of triazenes:

$$\log(1/C) = 0.10\cdot\log P - 0.04(\log P)^2 - 0.31\cdot\sigma^+ + 0.18\cdot MR_{2,6} + 0.39\cdot E_s - R + 4.12 \quad (36)$$
$$n = 61$$
$$r = 0.836$$
$$s = 0.19$$
$$\log P_o = 1.18.$$

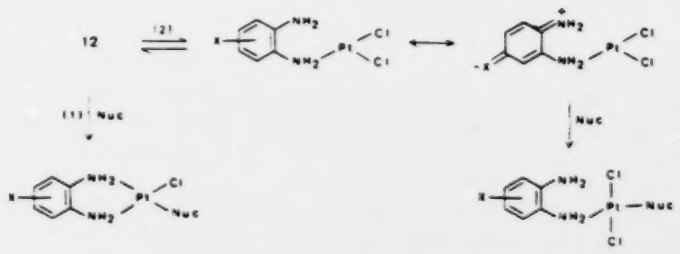
The quantity C in equation 36 is the moles/kg producing a 40% increase in life span in a test animal. The quantity $MR_{2,6}$ is a measure of the steric effect at positions 2 and 6 on the aromatic ring (ie, the positions ortho to the amino function). The crucial point to notice here is the lessened dependence of this activity on logP and σ^+.

According to Hansch et al [1979], this might allow the design of an optimal antitumor agent while minimizing the mutagenic properties. A more general conclusion to be derived from the mutagenic study is the expectation that, where there is microsomal activation, logP is likely to be important. Hansch further suggested, from these findings, that bacterial cells could be better suited for uncovering mutagenicity of hydrophilic compounds than mammalian test systems.

Recently, Hansch et al [1980] looked at the platinum compounds 12 with respect to mutagenicity in the test system *S typhimurium* TA92 with S9 activation. The activation system is not required for mutagenic activity but was used for consistency's sake in comparing these results with some of the previous studies. The following regression equation was derived:

$$\log 1/C = 2.23\ \Sigma(\sigma^-) + 5.78 \qquad (37)$$
$$n = 13$$
$$r = 0.98$$
$$s = 0.26.$$

Again, the quantity C is the molar concentration required to produce 30 mutations/100 million survivors and σ^- is the electronic parameter used when negative character is developed at the reaction center and delocalized by through-resonance. The facts that the coefficient for σ^- is fairly large and that Hammett's original σ does not give as high a correlation suggested a choice between two possible mechanistic pathways, shown in Scheme II.



Scheme II.

If the SN_2-like pathway 1 was operative, then it would be expected that the ρ value [the coefficient for $\Sigma(\sigma^-)$] in equation 37 would be much smaller: not as much charge would be developed in the transition state. In addition, Hammett's original σ should have been as useful as σ^-. Because the opposite is true, it may be assumed the mechanistic pathway is more SN_1-like in character as indicated in pathway 2.

Another interesting feature of this particular regression analysis is the lack of importance of logP. This is in spite of the broad range of π values examined: -0.04 to 2.13. According to Hansch, this is quite unusual for a biological QSAR study. He suggested that this is because bacteria contain no nucleus and therefore the DNA can be approached directly by even relatively hydrophilic agents (especially since, as was pointed out earlier, these compounds require no activation).

Triazenes were the subject of another study of structure and activity utilizing Hansch parameters and the MSA steric parameter [Hopfinger and Potenzone, 1982].

Wishnok and Archer [1976] looked at the relationship between the carcinogenicity of a set of N-nitrosoamines (13 in Fig. 1). The data were from the work of

Druckery et al [1967]. The carcinogenic index reported by Druckery et al was D_{50}, which is the dosage in mol/kg body weight required to cause tumors in 50% of the BD rats tested. Wishnok and Archer attempted correlation of $1/D_{50}$ with hexane-aqueous buffer partition coefficients and numbers of carbon atoms (N) and found only the numbers of carbons correlated:

$$\log(1/D_{50}) = 2.94 - 0.2N \quad (38)$$
$$n = 47$$
$$r = 0.64.$$

The authors speculated that this reflected steric, polarity, or electronic effects. Based on their equation, a carbon number of greater than about 14 carbons would result in inactivation.

In an extension of this earlier work, Wishnok et al [1978] considered a Hansch correlation between the same carcinogenic index and parameters for N-nitrosoamines. The parameters used were π (derived from hexane-aqueous buffer partition coefficients), Taft's inductive σ constant, plus some indicator variables for: 1) cyclic nitrosoamines with the amino nitrogen in the ring and acyclic nitrosoamines, 2) presence and location of OH^- functionality, aromaticity, and double bonds, 3) the number of Hs located alpha to the nitroso function and the number of nitroso groups. This resulted in the following regression equation:

$$\log(1/D_{50}) = 4.08 - 0.23\pi^2 + 0.64\pi - 0.25A - 0.72B - 0.70C \quad (39)$$
$$n = 24$$
$$s = 0.29$$
$$r^2 = 0.89$$

The authors noted that they attempted correlations with the number of Hs beta to and on the amino nitrogen with no success. The fact that only correlations were obtained with the alpha Hs was taken as evidence that these were quite important for activity.

Another study of nitrosoamines by Lijinsky's group [Singer et al, 1977] correlated relative mean life times (RML) and relative number of tumor-bearing animals (RTBA) but only if the nitrosoamines were acyclic:

$$\log(RML) = -0.262\ (\log P)^2 + 0.195\ \log P + 0.0112 \quad (40)$$
$$n = 6$$
$$r^2 = 0.99$$
$$s = 0.05$$

$$\log(RTBA) = -0.0813\ \log P + 0.09 \quad (41)$$
$$n = 6$$
$$r^2 = 0.94$$
$$s = 0.019.$$

However, the number of compounds studied was small.

Biagi et al [1983] found, via a structure-activity study of 5-nitroimidazoles (14 in Fig. 1) and mutagenicity, that reduction potentials gave only a very poor correlation. This is in spite of a number of studies reported in the Biagi paper indicating that this property is important and a recent report that correlated energies of the LUMO

of nitroarenes and mutagenicity [Klopman et al. 1984]. However, the variance in the log1/C as well as in the electroreduction potentials was low. There was a good correlation development in an equation that included the hydrophobicity parameter R_m (derived from chromatographic experiments) and a cross parameter ($MR^2 \times Hb$) that combines molar refractivity and the ability to hydrogen bond. The Niculescu-Duvaz et al [1981] study mentioned above in connection with MSD steric parameters also examined a carcinogenic database of heterocyclic aromatics.

This does not represent an exhaustive list of Hansch-type analyses of mutagenic or carcinogenic databases but is meant to allow one to gain a feel for this type of work.

Mutagenic and Carcinogenic Studies Utilizing SIMCA

There appear to be only a few studies utilizing SIMCA in the analysis of mutagenic and carcinogenic databases, such as the one by Dunn and Wold [1981]. In this study, the authors looked at carcinogenic activity of the substituted N-nitrosocompounds (13 in Fig. 1). There were a total of 61 such compounds; 11 were inactive, 50 active as carcinogens. The variables used in the initial principal component analysis were Rekker's f value, MR, σ^*, E_s, and the steric parameters L (length of the substituent) and B4 (width of substituent perpendicular to L and at widest point) of Verloop et al [1976]. The logP of the molecules was also used. This step failed to indicate any structure to the group as a whole. This indicates that the set does not represent a single class. The next step was to produce a class or classes independently showing some structure. The inactive compounds were assumed to have no structure and were no longer treated. The active compounds were divided into three somewhat arbitrary classes. The first division was accomplished by considering which of the active compounds required metabolic activation to express their carcinogenic properties and those that did not. The set requiring activation was further broken into two more sets: one that would be expected to yield a cation on activation and the other a diazoalkane. These groups did exhibit some structure; that is, their activity could be described by an equation like 21 within their class. In fact, for $-2 < u_1 < 0$ and $-2 < u_2 < 0$, there were two separated clusters of compounds. These clusters turned out to have some biological significance. One cluster represented compounds causing tumor formation of the esophagus, liver, and nasal turbinates; the other cluster of compounds were those compounds causing tumors only in the liver (all compounds were orally administered). In the case of those compounds causing esophageal cancer, σ^* was an important physicochemical factor, electron withdrawing or electron neutral substituents having greater activity.

Dunn and Wold [1978] also evaluated the carcinogenic potential of the 4-nitroquinoline-1-oxides (15 in Fig. 1). Again, Hansch-like variables were used. The noncarcinogens had no internal structure as a whole. Four components ($A=4$ in equation 21) were used to describe the variation in the active class. The 6-substituted compounds could also be broken into two separate classes, one active and one inactive. An identical set of ten variables could be used to describe the variance within each class.

The SIMCA method was also used to analyze class structure within 32 PAHs [Norden et al. 1978]. There were a mixture of theoretical and experimentally measured variables used in the analysis. The theoretical variables included the Pullman indices for bond localization and carbon localization energies for the K and L regions [Pullman and Pullman, 1955].

Mutagenic and Carcinogenic Studies Using ADAPT

The Jurs group has been one of the more active groups in the field of SAR studies on mutagens and carcinogens. One of the Jurs group's earliest studies [Chou and Jurs, 1979b] was that done on 153 compounds, 118 carcinogens and 35 noncarcinogens, containing the N-nitroso group. As discussed above, the ADAPT program is a pattern recognition method. The compounds are defined in a variable space, which is searched for a discriminant separating carcinogens from noncarcinogens. The variables belong to several different classes: topological, geometrical, physiological, and electronic. There were 15 descriptors found useful in discriminating between carcinogens and noncarcinogens in this database. These included six fragment descriptors (numbers of C atoms, N atoms, bonds, aromatic rings, double bonds, and basis rings), five molecular connectivity variables, one geometric variable, and one environment and two sigma charge descriptors. The ADAPT program has a number of options for the statistical pattern searching including quadratic Bayes, linear Bayes, K-nearest neighbor (K = 3), iterative least squares linear discriminants, simplex optimization, and the linear learning machine. In this particular instance, the linear learning machine and the iterative least squares routine produced the best separations between carcinogens and noncarcinogens. The overall predictive ability for the linear learning machine was evaluated at 91%. The evaluation was done by doing several training runs on the original database with ten compounds taken out (randomly except for preservation of the relative number of carcinogens to noncarcinogens). After each training run, the ten remaining compounds were evaluated using the descriptors generated. The average percentage correct over all such evaluations was the predictive ability.

Jurs demurs on generalizing his descriptors by saying that the descriptors are useful only in context. However, he indicates that the presence of the two electronic descriptors might indicate the importance of alpha-hydroxylation to the carcinogenic process. This had been implied, for instance, by the experimental work of Carrier et al [1978]. In addition, the ratio of large to intermediate principal moments was retained and suggests that molecular shape is crucial for carcinogenicity.

Another database treated by the ADAPT program consisted of aromatic amines of the general structures indicated in Figure 9 [Yuta and Jurs, 1981]. There were 157 compounds in total tested in rats by oral and injection administration routes. Tumors resulted in various organ sites including breast, ear duct, and liver. Less specific tumor sites were also indicated but were lumped into one category. The analysis was done on the compounds according to tumor site and administration protocol. Also, a mixed database was generated from combination of all results and analyzed by ADAPT. The same classes of descriptors and the same statistical evaluations were used as before. The substructures used are shown in Figure 10. The K-nearest neighbor was the least useful discriminant generator, being unable to classify 40–30% of the compounds. The linear discriminant generated by least squares produced the best separation of 90% correct. There were also some differences in the quality of the separations depending on which subset of the database was being examined. In general, the subsets of mixed biological end points were more difficult to separate. The subset consisting of mixed tumor sites and mixed administration routes was the single most difficult to fit.

Some of the more important descriptors in each subset of the database were the number of basis rings, number of oxygen atoms, molecular connectivity of substruc-

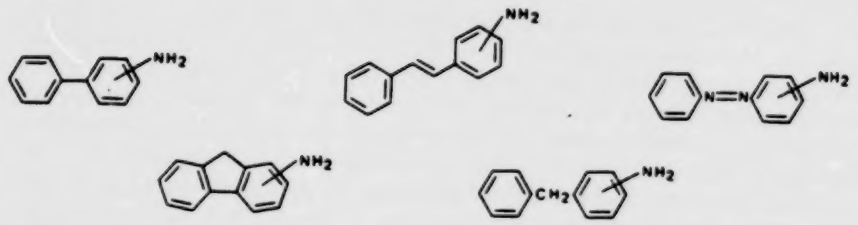


Fig. 9. Aromatic amines treated by ADAPT [Yuta and Jurs, 1981].

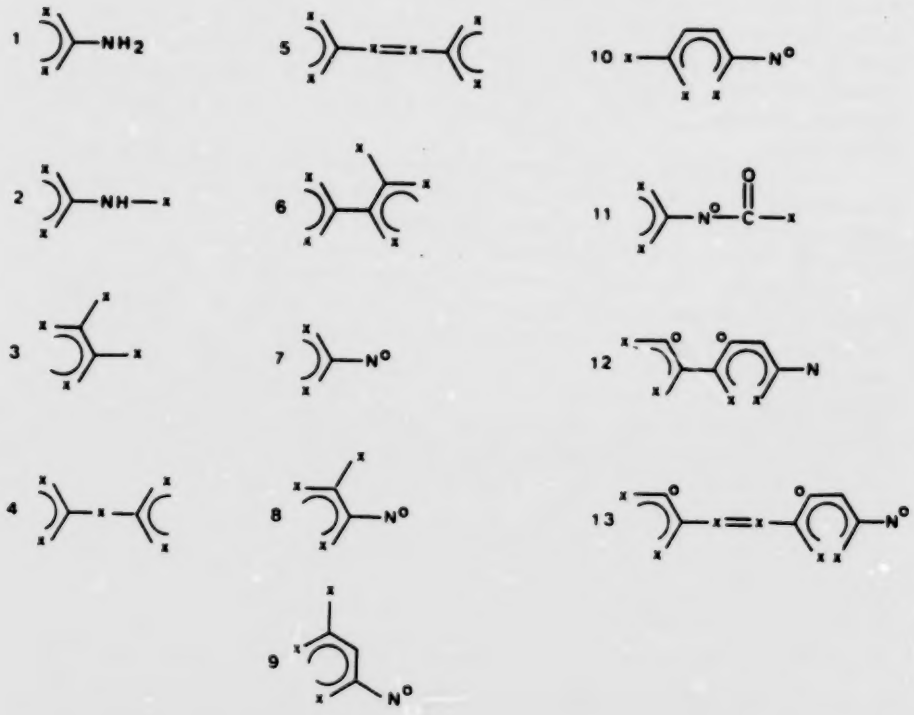


Fig. 10. Important substructures from ADAPT analysis of the aromatic amine database of Figure 9. The x's indicate an unspecified atom type but with the connectivity as shown; the o's indicate the specific atom type shown but with the connectivity of the atom unspecified.

ture 1 in Figure 10, and molecular connectivity environment for substructure 13. Jurs interprets the basis ring importance as suggesting that molecular volume is crucial.

The overall predictive ability, as developed above, was established at 85–90% for the mixed end point subset of the aromatic amine database.

ADAPT was also used to evaluate a chemically diverse database of 130 carcinogens and 79 noncarcinogens [Jurs et al. 1979]. The learning machine was the best classifier, correctly placing approximately 90% of the original total database. With 17 misclassified compounds removed, the learning machine required 26 descriptors to correctly place the remaining 192 compounds. The predictive ability derived from this reduced database was 85%. (This was evaluated as before by removal of ten compounds from the 192, analyzing the remaining 182, and predicting the placement of the excluded ten and repeating this process ten to fifteen times to obtain an average percentage correctly predicted). The predictive ability among carcinogens was 90%, for noncarcinogens 78%.

A portion of the Dipple [1976] database consisting of 200 polycyclic aromatics was treated with ADAPT also [Yuan and Jurs, 1980]. The descriptors utilized came from each of the descriptor categories. Among these were the substructure representing the bay region (recall that the substructures are entered by the researcher before the analysis begins), logP, descriptors representing the area of the molecules and their volume and shape, as well as the number of rings in the structure. Electronic descriptors included that of the average sigma charge of the substructure representing the bay region.

Jurs and Yuan were able to obtain a decision surface (discriminant) of 28 terms that correctly separated 191 (96%) of the 200 compounds into active and inactive classes. In predictive trials, the same decision surface correctly placed approximately 90% of the compounds as to whether they were carcinogens or noncarcinogens.

An extended version of the N-nitroso database previously reported by the Jurs groups was recently evaluated [Rose and Jurs, 1982]. It consisted of 150 compounds (112 carcinogens and 38 noncarcinogens) from the literature. The usual descriptors were used with the addition of one new descriptor, a symmetry parameter. The symmetry parameter is defined as the ratio of the number of uniquely connected atoms over the total number of connected atoms. There were six substructures chosen (Fig. 11) with the expectation that positions alpha and beta to the N-nitroso group would be especially important in biological alpha-hydroxylation of the N-nitroso compounds [Rose and Jurs, 1982].

With 15 descriptors, 93% (140/150) of the compounds were correctly placed as carcinogens or noncarcinogens. The percentage correctly expected on the basis of guessing can be calculated for comparison:

$$P = [(a/T)^2 + (b/T)^2]. \tag{42}$$

where a is the number of compounds in one category, b is the number of compounds in the other category, and T is the number of compounds total. The value P for this database distribution is 62%. The overall predictive ability has been evaluated by Rose and Jurs to be about 86%, with a slightly higher value for the predictive ability for the carcinogens alone of 90%.

For this database, Rose and Jurs also explained which descriptors accounted for how many of the correctly placed compounds. The most positive charge of substructures 3 and 5 accounted for approximately 13% of the observed separation. This descriptor was pointed out to be important not only as an indicator of reactivity but also as a determinant of transport properties. The molecular size and shape descriptors (basis rings, principal moments, and their ratios) accounted for nearly 20% of the separability. The four substructures representing the alpha and beta positions to the N-nitroso group were able to separate 97% of the data, again supporting the importance of these positions in activation to the ultimate carcinogenic species.

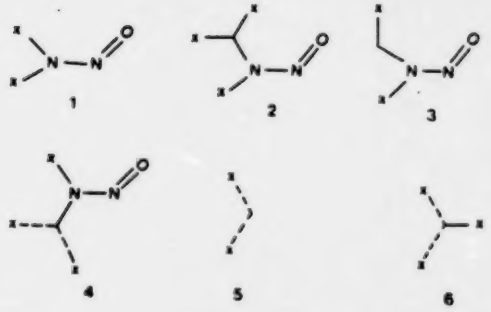


Fig. 11. Important substructures from ADAPT analysis of extended N-nitroso database of Rose and Jurs [1982].

CASE Analyses Related to Mutagenicity and Carcinogenicity

The CASE program has shown itself useful in the analysis of both carcinogenic and mutagenic data. One of the first CASE analyses of a mutagenic database looked at 53 nitroarenes tested in *S typhimurium* TA98 in the absence of S9 activation [Rosenkranz and Mermelstein, 1983]. There were 26 compounds classified as mutagens, 22 as nonmutagens, and five as marginal. The CASE analysis was able to classify correctly 89% of these compounds on the basis of structural information alone. There were identified structural fragments known to be involved in mutagenesis as well as some interesting new ones [Klopman and Rosenkranz, 1984]. Some of the fragments found to be associated with mutagenicity in that study are shown in Figure 12. Of particular interest are fragments 1 and 3. One indicates that there must be continuation of aromaticity para to the nitro substituent. On the other hand, if this position is unsubstituted, ie, a hydrogen is present, the resulting "biophobe" is actually deactivating.

The compounds that were not correctly classified in this database mainly consisted of compounds containing more than five rings. The implication is that, although these compounds may contain some of the same activating features as the smaller ring structures, their bulk prevents them from crossing cell membranes.

Another study looked at 56 compounds, all PAHs: 17 carcinogenic, 33 noncarcinogenic, and six only marginally carcinogenic or indeterminate [Klopman et al, 1985c]. This database was from the literature [Arcos and Argus, 1968]. The analysis identified three fragments having > 99% chance of being associated with the carcinogenicity (or lack of) for the training set (see Fig. 13). In spite of the fact that the program received no information other than the structures and activities of the compounds, these fragments represent essential features of the bay region, L region [Pullman and Pullman, 1955; Pullman, 1964], and pseudobay region [Perin-Roussel et al, 1984]. Of these compounds, 86% were correctly placed in their respective activity categories. Another PAH database examined resulted in essentially the same conclusions [Klopman, 1984].

Mitchell et al [1985] evaluated substituted polycyclic aromatic hydrocarbons that had been tested for both carcinogenic and mutagenic activity with the CASE program. Upon comparison of the results, the authors were able to conclude that there were few overlapping fragments. This seems to support previous observations

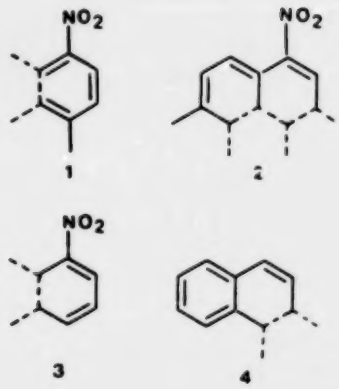


Fig. 12. Fragments (substructures) chosen by the CASE program after analysis of the aromatic nitro database. Fragments 1 and 2 were considered activating for mutagenicity in the *S typhimurium* TA98 strain (without S9 activation), and fragments 3 and 4 were inactivating.

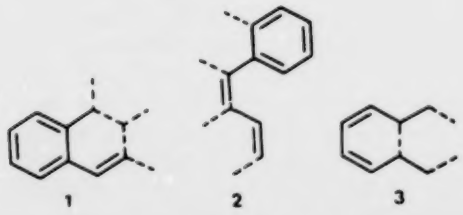


Fig. 13. Fragments chosen by the CASE program analysis of a PAH database. Fragment 1 illustrates the essential features of the bay region, fragment 2 represents those of the pseudobay region, and fragment 3 shows the essential features of the L region (see Fig. 8).

that, for instance, the bay region is important in carcinogenesis but much less so in mutagenesis.

Aromatic amines tested in the *S typhimurium* strains TA98 and TA100 with S9 activation have also been analyzed by CASE [Klopman et al. 1985b]. The compounds studied here included monocyclic and polycyclic aromatic amines, 80 tested in the TA98 strain and 107 tested in the TA100. From the CASE fragment analysis, two general comments are possible: 1) Substitution of the nitrogen functionality by alkyl groups tended to be a deactivating feature in both the TA98 and TA100 databases and 2) The deactivating fragment $-CH_2CH_2OH$ decreased activity whether attached to a nitrogen or not. Within the broad classification of mutagen or nonmutagen, CASE placed 88% of the compounds of the TA98 database and 84% of the compounds in the TA100 database. However, since the data were input in a semi-quantitative form, a fragment-based QSAR was possible. The results of this analysis are shown in the "truth matrix" summary of Table IVa and b.

TABLE IVa. Summary of CASE Calculated vs Experimental Results for Amines Tested Against TA98 with S9

Experimental	Calculated					
	–	+	++	+++	++++	Total
–	28	6	0	0	0	34
+	5	2	1	0	0	8
++	0	9	2	2	1	14
+++	1	5	1	4	1	12
++++	0	0	0	3	14	17
Total	34	22	4	9	16	85

TABLE IVb. Summary of CASE Calculated vs Experimental Results for Amines Tested Against TA100 with S9

Experimental	Calculated					
	–	+	++	+++	++++	Total
–	39	1	0	2	1	43
+	6	0	0	2	0	8
++	5	0	0	2	3	10
+++	0	0	0	5	6	11
++++	2	0	0	5	28	35
Total	52	1	0	16	38	107

Finally, for a set of 19 compounds tested in the TA98 strain, 84% were placed correctly within the broad classification of mutagens or nonmutagens.

A pesticide database has also been analyzed by CASE [Klopman et al, 1985a]. The 54 compounds examined had been tested in five short-term assays: the *S typhimurium mutagenicity* assay (SAL), a specific gene mutation assay using mouse lymphoma L5178Y cells (L5T), prokaryotic DNA repair assays using, respectively, pairs of *Escherichia coli* polA$^+$ and polA$_1^-$ (REP) and *Bacillus subtilis* rec$^+$ and rec$^-$ (REW), and recombinogenicity in the yeast *Saccharomyces cerevisiae* (YE3). The test results were only qualitative. There were several points of interest in the CASE analysis. The most important fragments derived from the analysis of the five tests are depicted in Figure 14. The three fragments 1, 2, and 4 were found to be activating in compounds tested in the tests SAL, YE3, and L5T. The fragments 5 and 6 were found to be inactivating in compounds in the same tests. These facts indicate the possibility of a mechanistic relationship between mutagenicity in a prokaryote (*S typhimurium*), recombinogenicity (ie, DNA damage) in a eukaryote (*S cervisiae*), and mutagenicity in mammalian cells (L5T). There was not much overlap between fragments obtained on the analysis and comparison of the REP and REW databases. Fragments 2, 7, and 8 were activating and fragment 9 was inactivating in the REW database. For the REP database, there were three activating fragments: 2, 10, and 11. The differences between the REP and REW results were imputed to be the result of the differences in cell wall structure between the two strains of organisms (the *E coli* being a gram-negative species and the *B subtilis* being gram-positive). Gram-negative organisms have a cell wall that facilitates passage of lipophilic molecules, whereas

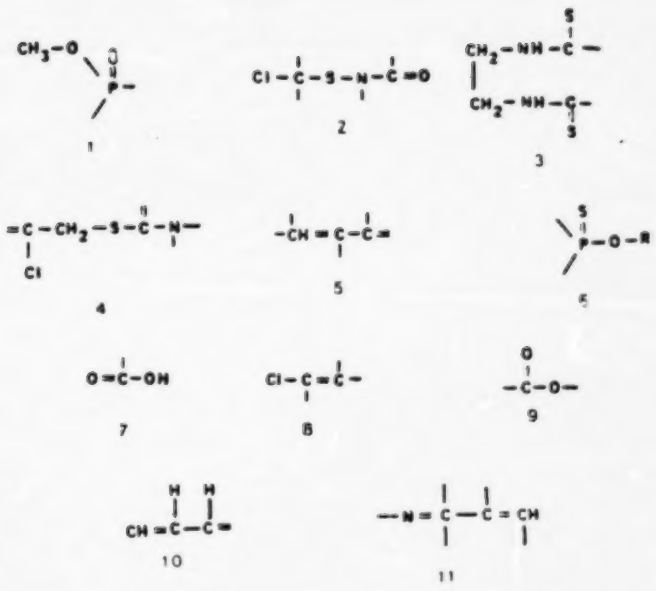


Fig. 14. CASE fragments important in the SAL, L5T, YE3, REW and REP tests for mutagenicity.

the gram-positive organisms have a cell wall that facilitates passage of hydrophilic species.

A comparison across the battery of the five tests produced fragment 2 as activating in all the tests. There was no overlap (other than fragment 2) between the REP and REW and the other three tests consistent with a classification scheme by Waters et al [1982].

The Jurs database of carcinogens [Yuan and Jurs, 1980] was submitted for CASE analysis; these data were on hand to a large extent, and it was thought a direct comparison of these two methods might be informative. Jurs was able to classify correctly 95% of the original 200 polycyclic aromatic hydrocarbons using 28 descriptors. The latest version of CASE treating the same 200 compounds in a semi-quantitative fashion correctly classified 89% [Klopman and Frierson, 1986]. This is somewhat lower than the ADAPT program of Jurs. However, this was achieved with less than half as many descriptors as the ADAPT program. The discrepancy might be due to the fact that the study of PAHs as carcinogens is a very well developed area of SAR studies. The ADAPT program is able to take advantage of this knowledge in its design to accept just about any descriptor one might wish to include. In this particular case, that included information about the bay region and charge densities around the periphery of the PAHs, both of which are known to have been correlated to carcinogenicity [Yuan and Jurs, 1980]. On the other hand, the CASE program designs its classification scheme without any input or knowledge from the researcher except the structure of the molecule and its biological activity. Clearly, this might be a disadvantage in cases that have been well studied from an SAR point of view but might be a distinct advantage in those cases for which little is known about what makes a structure biologically active.

Enslein Method Applied to Carcinogenicity and Mutagenicity

Enslein and Craig [1982] used their methods on a database collected from the International Agency for Research on Cancer (IARC) Monographs "Evaluation of Carcinogenic Risk of Chemicals to Humans" (volumes 1–17) [IARC, 1972–1978]. They used 343 of the compounds available from this database, 120 noncarcinogens and 223 carcinogens. (Although IARC has six categories of activity, Enslein and Craig combined noncarcinogens with indefinite human carcinogens categories and also combined the definite human and animal carcinogens categories into the two above.) The database was chemically quite diverse containing a variety of chemical classes. The descriptors were substructural fragments obtained from WLN codes, the molecular weight, and the molar refractivity. Some additional fragment descriptors (keys) were employed that were derived from known carcinogens or were required to code special functionalities.

The discriminant equation utilizing some 79 descriptors (all but one, the molecular weight, of a substructural type) correctly classified between 87% and 91% of the carcinogens and 78% and 80% of the noncarcinogens. The indeterminancy comes about because the developed discriminant equation defines only a probability of a compound being a carcinogen, resulting in regions about 50% that are undefinable in terms of whether a compound is a carcinogen or not. Enslein and Craig did not attempt in this study to evaluate a test set of compounds not in their original database. However, they pointed out that compounds not possessing features of the training set were likely to be misclassified. For this reason also some compounds were "strongly misclassifed," meaning that they were given high probabilities of belonging to a class by the discriminant equation but that this predicted class was incorrect. They were eight such compounds. One, 2,4-diaminotoluene, was misclassified in every model that the Enslein group tried. Enslein and Craig suggest such a compound needs to be reexamined experimentally.

The large numbers of relatively unranked descriptors utilized in this study do not lend themselves to a mechanistic interpretation. Indeed, Enslein and Craig recommend against trying to interpret their results in this way. The discriminant equation derived is for purposes of classification only and only within the chemical space represented by the compounds of the database.

The Enslein group has also looked at a relatively large database containing mutagens and nonmutagens of 532 compounds [Enslein et al, 1983b]. These compounds came from two sources: the Environmental Mutagenesis Information Center (EMIC; Oak Ridge, Tennessee) and the National Toxicological Testing Program (NTP). Because these sources report mutagenic activity in a number of different tests, some sort of criteria had to be developed to define the active and inactive classes for use in the discriminant analysis. The criteria used defined mutagens as those compounds that had at least one positive response (with or without S9 activation) in the *S typhimurium* strains tested (TA98, TA100, TA1535, TA1537, and TA1538). A compound was defined as a nonmutagen when it had been tested in the three strains TA1535, TA1537, and TA1538 and been found negative in each case. This gave 301 mutagens and 231 nonmutagens. Of these, 264 of the mutagens and 208 of the nonmutagens were used in developing the discriminant equation, and 37 mutagens and 23 nonmutagens were held back for evaluating the model.

The substructural fragments were the CROSSBOW WLN-based keys discussed above [Eakin et al, 1974] plus the additional keys developed by Enslein to better

represent his databases and to provide substructure keys thought to be related to carcinogenic activity, which, however, were not further identified [Enslein et al, 1983b]. Also included in this database were the molecular weights as descriptors.

The model-building routine was the same as previously described. The discriminant equation ultimately derived was able to classify correctly 86% of the mutagens in the database. When applied to the test set mutagens, the classification was about 80% correct. Although there is listed a set of 20 descriptors described as most important, Enslein again cautions against interpretation; apparently, their ranking could change greatly with changes in the database. The fluctuations in descriptor importance could be damped to near zero by the inclusion of approximately 1,500 compounds in the database. This number was estimated from results in the rat oral LD_{50} database treated by this method [Enslein et al, 1983a].

One systematic misclassification was the class of mutagens known as intercalators. This is easily explained on consideration of the types of descriptors used by the Enslein method. This particular class of mutagens is probably strongly dependent on topological features and possibly on the ability to hydrogen-bond [Enslein et al, 1983b] of the intercalating agents. The coded substructures are expressed in two dimensions and therefore would not be expected to provide much information of a topological nature.

FUTURE DEVELOPMENTS

It is essential that computer-aided storage, retrieval, and analysis of the tremendous volume of data on carcinogens/mutagens and potential carcinogens/mutagens become available to researchers and decision-makers in the environmental science field. Because of the volume of data being generated in research labs throughout the world, there are no longer (if there ever were) single individuals or groups of individuals whose judgments of the carcinogenic and genotoxic effects of new, potentially harmful substances could possibly take into account the whole body of experimental knowledge. Most of the connectivity approaches that have been discussed are capable, in principle, of doing this. In addition, they do this in a critical fashion; that is, they were designed to have built in the analytical side of the judgement-making process of the human expert. That is not to say that computer judgement surpasses or is even equal to that of the human expert; in many ways artificial intelligence will remain artificial. However, these "expert" systems have demonstrated that they have an acceptably high degree of accuracy in several important databases. There is still room for improvement and therefore continued need for the support and development of these methods.

One shortcoming both approaches (extrathermodynamic and connectivity) presently have is the inability to represent truly the three-dimensionality of the compounds they analyze. This is a major shortcoming; biological specificity is predicated on the shape and points of interaction of biomolecules, including the handedness of the molecules. Many research groups (including our own) are working to overcome this deficiency.

ACKNOWLEDGMENTS

We wish to thank the Office of Naval Research, the National Institute of Environmental Health Sciences, and the U.S. Environmental Protection Agency for the support of much of this work.

REFERENCES

Arcos JC, Argus MF (1968): Molecular geometry and carcinogenic activity of aromatic compounds: New perspectives. Adv Cancer Res 11:305–471.

Badger GM (1948): The carcinogenic hydrocarbons: Chemical constitution and carcinogenic activity. Br J Cancer 2:309–350.

Badger GM (1949): The relative reactivity of aromatic double bonds. J Chem Soc 456–463.

Badger GM (1950): The relative reactivity of aromatic double bonds. Part III. The relation between double-bond character and the velocity of addition of osmium tetroxide. J Chem Soc 1809–1814.

Badger GM (1954): Chemical constitution and carcinogenic activity. Adv Cancer Res 2:73–127.

Badger GM, Lynn KR (1950): The relative reactivity of aromatic double bonds. Part II. Addition of osmium tetroxide to substituted 1:2-benzanthracenes. J Chem Soc 1726–1729.

Badger, GM, Pearce RS, Pettit R (1952): Substituted anthracene derivatives. Part V. The conjugating powers of the substitution positions in 1:2-benzanthracene. J Chem Soc 1112–1116.

Bergmann F (1942): The mechanism of tumor production by chemical agents. Cancer Res 2:660–663.

Bevington PR (1969): "Data Reduction and Error Analysis for the Physical Sciences." New York: McGraw-Hill Book Co., pp 164–203.

Biagi GL, Barbaro AM, Guerra MC, Forti GC, Aicardi G, Borea PA (1983): Quantitative relationship between structure and mutagenic activity in a series of 5-nitroimidazoles. Teratogen Carcinogen Mutagen 3:429–438.

Boyland E (1950): The biochemistry of abnormalities in cell division. Biochim Biophys Acta 4:293–300.

Brown HC, Stock LM (1963): A quantitative treatment of directive effects in aromatic substitution. Adv Phys Organic Chem 1:35–154.

Burkert U, Allinger NL (1982): "Molecular Mechanics." Washington, DC: American Chemical Society.

Carrier AM, Wiesak M, Molaveille C, Bartsek H (1978): High mutagenicity of N-(α-aryloxy)alkyl-N-alkylnitrosoamines in *S. typhimurium*: Model compounds for metabolically activated N,N-dialkylnitrosoamines. Mutat Res 49:187–194.

Chou JT, Jurs PC (1979a): Computer-assisted computation of partition coefficients from molecular structures using fragment constants. J Chem Inf Comp Sci 19:172–178.

Chou JT, Jurs PC (1979b): Computer-assisted structure-activity studies of chemical carcinogens. An N-nitroso compound data set. J Med Chem 22:792–797.

Chu KC, Feldman RJ, Shapiro MB, Hazard GF, Geran RI (1975): Pattern-recognition and structure-activity relationship studies. Computer-assisted prediction of antitumor activity in structurally diverse drugs in an experimental mouse brain tumor system. J Med Chem 18:539–545.

Clark J, Perrin DD (1964): Predictions of the strengths of organic bases. Quart Rev 18:295–320.

Collander R (1954): The permeability of nitella cells to nonelectrolytes. Physiol Plantarum 7:420–445.

Cook JW, Schowenthal R (1948): Oxidation of carcinogenic hydrocarbons. J Chem Soc 170–173.

Coulson CA (1953): Electronic configuration and carcinogenesis. Adv Cancer Res 1:1–56.

Craig PN (1971): Interdependence between physical parameters and selection of substituent groups for correlation studies. J Med Chem 14:680–684. 1251.

Daudel P (1950): Application de la methode des états de spin à l'étude des complexes intermediaires. Cr Acad Sci 230:99–100.

Daudel P, Daudel R (1950): Applications of wave mechanics to the study of the mode of action of carcinogenic substances on tissues. Biol Med (Paris) 39:201–236.

Daudel R (1948): Sur les applications de la mécanique ondulatoire à l'étude des phénomènes de cancerisation. Bull Cancer 35:110–124.

Del Re G (1948): A simple MO-LCAO method for method for the calculation of charge distributions in saturated organic molecules. J Chem Soc 4031–4040.

Dewar MJS, Dougherty RC (1975): "The PMO Theory of Organic Chemistry." New York: Plenum Press, pp 212–220.

Dewar MJS, Grisdale PJ (1962): Substituent effects I. An introduction and Substituent effects IV. A quantitative theory. J Am Chem Soc 84:3539–3541, 3548–3553.

Dipple A (1976): Polynuclear aromatic carcinogens. In Searle CE (ed): "Chemical carcinogens." Washington, DC: American Chemical Society, pp 245–314.

Dixon WJ (ed) (1977): "BMD Biomedical Computer Programs." Berkeley: Univ of California Press, p 233, (BMDP7M, Stepwise Discriminant Analysis).

Druckery A, Preussman R, Ivankovic S, Schmahl D (1967): Organotrope carcinogene wirkungen bei 65 verscheiden and BD-ratten. Z. Krebsforsch 69:103–201.

Dunn WJ III, Greenberg MJ, Callejas SS (1976): Use of cluster analysis in the development of structure-activity relationships of antitumor triazenes. J Med Chem 19:1299–1301.

Dunn WJ III, Wold S (1978): A structure-carcinogenicity study of 4-nitroquinoline-1-oxides using the SIMCA method of pattern recognition. J Med Chem 21:1001–1007.

Dunn WJ III, Wold S (1980): Relationship between chemical structure and biological activity modeled by SIMCA pattern recognition. Biorg Chem 9:505.

Dunn WJ III, Wold S (1981): The carcinogenicity of N-nitroso compounds: A SIMCA pattern-recognition study. Biorg Chem 10:29–45.

Eakin DR, Hyde E, Parker G (1974): Use of computers with chemical structural information—ICI (Imperial Chemical Industries, Ltd.) CROSSBOW system. Pesticide Sci 5:319–326.

Enslein K, Craig PN (1978): A toxicity estimation model. J Environ Pathol Toxicol 2:115–132.

Enslein K, Craig PN (1982): Carcinogenesis: A predictive structure-activity model. J Toxicol Environ Health 10:521–530.

Enslein K, Lander TR, Strange JR (1983a): Teratogenesis: A statistical structure-activity model. Teratogen Carcinogen Mutagen 3:429–438.

Enslein K, Lander TR, Tomb ME, Landis WG (1983b): Mutagenicity (Ames): Structure-activity model. Teratogen Carcinogen Mutagen 3:503–513.

Exxner O (1978): A critical compilation of substituent constants. In Chapman NB, Shorter J (eds): "Correlation Analysis in Chemistry." New York: Plenum Press, pp. 439–540.

Flynn GL, Yalkowsky SH (1972): Correlation and prediction of transport across membranes I: Influence of alkyl chain length in flux determining properties of barrier and diffusant. J Pharm Sci 61:838–852.

Free SM, Wilson JW (1964): A mathematical contribution to structure-activity studies. J Med Chem 7:395–399.

Fukui K, Nagata C, Imamura A (1960): Electronic structure and nicotine-like stimulant activity in choline. Science 132:87–88.

Gammans RE, Sehon RD, Anders MW, Hanna PE (1977): Microsomal N-hydroxylation of trans-4'-halo-4-acetamidostilbenes. Drug Metab Dispos 5:310–316.

Greenwood HH (1951): The reactivity of aromatic bonds, with reference to carcinogenic compounds of 1:2 benzanthracene. Br J Cancer 5:441–457.

Grunwald GL, Creese MW, Walters DE (1979): Applications of CNDO/2 calculations and X-ray crystallographic analysis to conformationally defined analogs of methamphetamine. In Olson EC, Christoffersen RE (eds): "Computer-Assisted Drug Design." ACS Symposium Series 112. Washington, DC: American Chemical Society, pp 439–487.

Haddow A, Harris RJC, Kon GAR, Roe EMF (1948): The growth inhibitory and carcinogenic properties of 4-aminostilbene and derivatives. Trans R Soc London, 241A:147–195.

Haddow A, Kon GAR (1947): Chemistry of carcinogenic compounds. Br Med Bull 4:314–326.

Hall DJ, Khanna D (1977): The ISODATA method. In Enslein K, Wolf HS, Ralston A (eds): "Statistical Methods for Digital Computers." New York: Wiley.

Hall LH, Kier LB (1977): Structure-activity studies using valence molecular connectivity. J Pharm Sci 66:642–644.

Hammett LP (1940): "Physical Organic Chemistry." New York: McGraw Hill.

Hansch C (1976): On the structure of medicinal chemistry. J Med Chem 19:1–6.

Hansch C (1978): Recent advances in biochemical QSAR. In Chapman, NB, Shorter J (eds): "Correlation Analysis in Chemistry." New York: Plenum Press, pp.397–438.

Hansch C (1971): Quantitative structure-activity relationships in drug design. In Ariéns EJ (ed): "Drug Design." New York: Academic Press, pp 271–337.

Hansch C, Fujita T (1964): ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626.

Hansch C, Leo A, Unger S, Kim KH, Nikaitani D, Lien EJ (1973): "Aromatic" substituent constants for structure-activity correlation. J Med Chem 16:1207–1216.

Hansch C, Muir RM, Fujita T, Maloney PP, Geiger CF, Streich, MJ (1963): The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc 85:2817–2824.

Hansch C, Muir RM, Fujita T, Maloney PP, Geiger CF, Streich MJ (1963): The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc 85:2817–2824.

Hansch C, Unger SH (1973): Strategy in drug design. Cluster analysis as an aid in selection of substituents. J Med Chem 16:1217–1222.

Hansch C, Venger BH, Hathaway GJ, Amrein YU (1979): Ames test of 1-(x-phenyl)-3,3-dialkyltriazenes. A quantitative structure-activity study. J Med Chem 22:473–476.

Hansch C, Venger BH, Panthananickal A (1980): Mutagenicity of substituted (o-phenyldiamine)platinum dichlorides in the Ames test. A quantitative structure activity study. J Med Chem 23:459–461.

Higuchi T, Davis SS (1970): Thermodynamic analysis of structure activity relationships of drugs: Prediction of optimal structure. J Pharm Sci 59:1376–1383.

Hirsch J (1974): "Concepts in Theoretical Organic Chemistry." Boston: Allyn and Bacon, Inc., pp 108–113, 128–132.

Hodes GF, Hazard R, Geran I, Richman S (1977): A statistical-heuristic method for automated selection of drugs. J Med Chem 20:469–475.

Hopfinger AJ, Battershell RD (1976): Application of SCAP to drug design I. Prediction of octanol-water partition coefficient using solvent dependent conformational analysis. J Med Chem 19:569–573.

Hopfinger AJ, Potenzone R Jr (1982): Ames test and antitumor activity of 1-(1-phenyl)-3,3-dialkyltriazenes. Mol Pharmacol 21:187–195.

IARC (1972–1978): Monographs on the Evaluation of Carcinogenic Risks of Chemicals to Man, Vols. 1–17. International Agency for Research on Cancer, Lyon.

Ingold CK (1930): Mechanism of, and constitutional factors controlling, the hydrolysis of carboxylic acid esters I. The constitutional significance of hydrolytic stability maxima. J Chem Soc 1032–1039.

Jerina DM, Lehr RE (1977): The bay-region theory: A quantum mechanical approach to aromatic hydrocarbon-induced carcinogenicity, in Ullrich V, Roots I, Hildebrant, AG, Estabrook RW, Conney AH (eds): "Microsomes and drug oxidations." Oxford: Pergammon Press, p 709.

Jones RN (1940): The ultraviolet absorption spectra of some derivatives of 1,2-benzanthracene. J Am Chem Soc 62:148–152.

Jurs PC, Chou JT, Yuan M (1978): Studies of chemical structure-biological activity relations using pattern recognition, in Olson EC, Christofferson RE (eds): "Computer-Assisted Drug Design." ACS Symposium Series 112. Washington, DC: American Chemical Society, pp 103–129.

Jurs PC, Chou JT, Yuan M (1979): Computer-assisted structure-activity studies of chemical carcinogens. A heterogenous data set. J Med Chem 22:476–483.

Jurs PC, Hasan MN, Henry DR, Stouch TR, Whalen-Pederson EK (1983): Computer-assisted studies of molecular structure and carcinogenic activity. Fund Appl Toxicol 3:343–349.

Jurs PC, Kowalski BR, Isenhour TL (1969): Computerized learning machines applied to chemical problems: Molecular formulas determined from high resolution mass spectrometry. Anal Chem 41:21–27.

Kier LB, Hall LH (1976a): Molecular connectivity VII: Specific treatment of heteroatoms. J Pharm Sci 65:1806–1809.

Kier LB, Hall LH (1976b): "Molecular Connectivity in Chemistry and Drug Research." New York: Academic Press.

Kier LB, Simons RJ, Hall LH (1978): Structure activity studies on mutagenicity of nitrosamines using molecular connectivity. J Pharm Sci 67:725–726.

Klopman G (1984): Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J Am Chem Soc 106:7315–7321.

Klopman G, Contreras R, Rosenkranz HS, Waters MD (1985a) Structure-genotoxic activity relationship of pesticides: Comparison between the results of several short-term assays. Mutat Res 147:343–356.

Klopman G, Frierson MR (1986): CASE analysis of Dipple database of 200 PAH's. In preparation.

Klopman G, Frierson MR, Rosenkranz HS (1985b): Computer analysis of toxicological databases: Mutagenicity of aromatic amines in Salmonella tester strains. Environ Mutagen 7:625–653.

Klopman G, Iroff L (1981): Correlation of partition coefficients by the charge density method. J Comput Chem 2:157–160.

Klopman G, Kalos A (1985): Causality in structure-activity studies. J Comput Chem 6:492–506.

Klopman G, Namboodiri K (1985): A simple method of computing the partition coefficient. J Comput Chem 6:28–38.

Klopman G, Namboodiri K, Kalos A (1985c): Computer automated evaluation and prediction of the

Iball index of carcinogenicity of polycylic aromatic hydrocarbons. In Rein R (ed): "Molecular Basis of Cancer. Part A: Macromolecular Structure, Carcinogens and Oncogenes." New York: Alan R. Liss, Inc., pp 287–298.

Klopman G, Rosenkranz HS (1984): Structural requirements for the mutagenicity of environmental nitroarenes. Mutat Res 126:227–238.

Klopman GD, Tonucci DA, Holloway M, Rosenkranz HS (1984): Relationship between polarographic potential and mutagaenicity of nitroarenes. Mutat Res 126:139–144

Kubinyi H (1977): Quantitative-structure activity relationships 7. The bi-linear model, a new model for non-linear dependence of biological activity on hydrophobic character. J Med Chem 20:625–629.

Leo A (1985): Pomona College Medicinal Chemistry Project.

Leo A, Hansch C, Elkins D (1971): Partition coefficients and their uses. Chem Rev 71:525–616.

Leo A, Jow PYC, Silipo C, Hansch C (1975): Correlation of hydrophobic constants (log P) from pi and f constants. J Med Chem 18:865–868.

Lewi PJ (1980): Multivariate data-analysis in structure-activity relationships. In Ariens EJ (ed): "Drug Design." New York: Academic Press, pp 307–342.

Lowry TH, Richardson KS (1976): "Mechanism and Theory in Organic Chemistry." New York: Harper and Row Publishers, pp 62–71.

Martin YC (1981): A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. J Med Chem 24:229–237.

Martin YC (1978): "Quantitative Drug Design. A Critical Introduction." New York: Marcel Dekker.

Martin YC, Hansch C (1971): Influence of hydrophobic character on the relative rate of oxidation of drugs by rat liver microsomes. J Med Chem 14:777–779.

Martin YC, Holland JB, Jarboe CH, Potnikoff N (1974): Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans. J Med Chem 17:409–413.

Marquardt DW, Snee RD (1975): Ridge regression in practice. Am Stat 29:3–20.

Meyer H (1899): Zur theorie der alkoholnarkose. Erste mitteilung. Welche eigenschaft der anasthetica bedingt ihre narkotische wirkung? Arch Exp Pathol Pharmakol 42:110–118.

Meyer J, van der Waterbeemd H, Testa B (1982): Eur J Med Chem 17:17.

Morgan SL, Deming SN (1974): Simplex optimization of analytical chemical methods. Anal Chem 46:1170–1181.

Miller EC, Miller JA, Sandin RB, Brown RK (1949): The carcinogenic activities of 2-acetylaminofluorene in the rat. Cancer Res 9:504–509.

Mitchell CS, Klopman G, Rosenkranz HS (1985): Computer automated evaluation of mutagenicity and carcinogenicity of selected polycyclic aromatic hydrocarbons. In "Polycyclic Aromatic Hydrocarbons. Ninth International Symposium." Columbus, OH: Batelle Press, in press.

Murray WJ (1977): Molecular connectivity and steric parameters. J Pharm Sci 66:1352–1354.

Newman MS, Kosak AI (1949): Orientation of benzo[c]phenanthrene. J Organic Chem 14:375–376.

Niculescu-Duvaz T, Tugulea M, Croisy A, Jacquigon PC (1981): A quantitative structure-activity analysis of the mutagenic and carcinogenic action of 43 structurally related heterocyclic compounds. Carcinogenesis 2:269–275.

Norden B, Edlund U, Wold S (1978): Carcinogenicity of polycyclic aromatic hydrocarbons, studied by SIMCA pattern recognition. Acta Chem Scand Series B 32:602–608.

Orchin M (1951): Steric effects in complex compound formation. J Organic Chem 16:1165–1169.

Osman R, Basch H (1984): On the mechanism of action of superoxide dismutase: A theoretical study. J Am Chem Soc 106:5710–5714.

Osman R, Weinstein H, Green JP (1979): Parameters and methods in quantitative structure activity relationships. In Olson EC, Christoffersen RE (eds): "Computer-Assisted Drug Design." ACS Symposium Series 112. Washington DC: American Chemical Society, pp. 71–77.

Overton E (1901): "Studien uber die Narkose." Jena, Germany: Fischer.

Palajda M, Rosenkranz HS (1985): Assembly and preliminary analysis of a genotoxic data base for predicting carcinogens. Mutat Res 153:79–134.

Perin-Roussel O, Ekert B, Barat N, Zaldja F (1984): DNA-protein crosslinks induced by exposure of cultured mouse fibroblasts to dobenzo[a,e]fluoranthene and its bay and pseudo-bay region dihydrodiols. Carcinogenesis 5:379–383.

Pet-Edwards J, Rosenkranz HS, Chankong V, Haimes YY (1985a): Cluster-analysis in predicting the carcinogenicity of chemicals using short-term assays. Mutat Res 153:167–185.

Pet-Edwards J, Chankong V, Rosenkranz HS, Haimes YY (1985b): Application of the carcinogenicity prediction and battery selection (CPBS) method to the Gene-Tox data base. Mutat Res 153:187–200.

Pullman A (1947): Study of the electronic configuration of organic molecules. Special study of cancerigenic hydrocarbons. Ann Chim 2:5–71.

Pullman A (1948): Electronic structure and cancerigenic power of the derivatives of 4-aminostilbene. CR Acad Sci 226:486–488.

Pullman A, Pullman B (1946): Rev Sci 84:145.

Pullman A, Pullman B (1948): A quantum mechanical study of the electronic structure of carcinogenic molecules. Acta Unio-Intern Contra Cancrum 6:57–63.

Pullman A, Pullman B (1955): Electronic structure and carcinogenic activity of some aromatic molecules. Adv Cancer Res 3:117.

Pullman B (1946): Effects of substituents on electron distributions: power of hyperconjugation of methyl, equilibrium of the enol-ketone forms. CR Acad Sci 222:1396–1398.

Pullman B (1948): Effect of substituents on the electronic structure of organic molecules. Bull Soc Chim France 15:533.

Pullman B (1964): Electronic aspects of the interactions between the carcinogens and possible cellular sites of their activity. J Cell Comp Physiol 64 [Suppl 1]:91–109.

Randic M (1975): On characterization of molecular bonding. J Am Chem Soc 97:6609–6615.

Rekker RF (1977): "The Hydrophobic Fragmental Constant, Pharmacochemistry Library." New York: Elsevier, Vol 1.

Rekker RF, deKort HM (1979): The hydrophobic fragmental constant: An extension to a 1000 point data set. Eur J Med Chem-Chim Ther 14:479–488.

Ritter GL, Lowry SR, Wilkins CL, Isenhour TL (1975): Simplex pattern recognition. Anal Chem 47:1951–1956.

Robinson R (1946): Aspects of chemistry related to medicine. Br Med J 1:943–946.

Rose SL, Jurs PC (1982): Computer-assisted studies of structure-activity relationships of N-nitroso compounds using pattern recognition. J Med Chem 25:769–776.

Rosenkranz HS, Mermelstein R (1983): Mutagenicity and genotoxicity of nitroarenes: All nitro-containing chemicals were not created equal. Mutat Res 114:217–267.

Sandin RB, Melby R, Hay AS, Jones RN, Miller EC, Miller JA (1952): Ultraviolet spectra and carcinogenic activities of some fluorene and biphenyl derivatives. J Am Chem Soc 74:5073–5075.

Schmidt O (1939a): Mechanism of the excitation process in cancerous and healthy cells. Z Phys Chem B44:194–202.

Schmidt O (1939b): Density distribution of B-electrons. II. Characterization of simple and cancer-producing hydrocarbons through the density distribution of certain valence electrons (B-electrons). Z Phys Chem B42:83–110.

Service J (1972): "A User's Guide to the Statistical Analysis System (SAS)." Raleigh, NC: N.C. State Univ, p 190–195.

Shear MJ, Leiter JJ (1941): Studies in carcinogenesis XVI. Production of subcutaneous tumors in mice by miscellaneous polycyclic compounds. J Natl Cancer Inst 2:241–258.

Siegel S (1956): "Non-parametric statistics for the Behavioral Sciences." New York: McGraw-Hill, pp 127–136.

Simon Z, Badilescu I, Racovitan T (1977): Mapping of dihydrofolate reductase receptor site by correlation with minimal topological differences. J Theor Biol 66:485–495.

Singer GM, Taylor HW, Lijinsky WJ (1977): Liposolubility as an aspect of nitrosoamine carcinogenicity: Quantitative correlations and qualitative observations. Chem Biol Interact 19:133–142.

Sjostrom M, Wold S (1974): Statistical analysis of the Hammett equation II. A unified inductive sigma scale. Chem Script 6:114–121.

Sjostrom M, Wold S (1976): Statistical analysis of the Hammett equation IV. An expanded sigma zero scale. Chem Scripta 9:200–210.

Stuper AJ, Brugger WE, Jurs PC (1977): A computer system for structure-activity studies using chemical structure information handling and pattern recognition techniques. In Kowalski BR (ed): "Chemometrics: Theory and Application." ACS Symposium Series 52. Washington, DC: American Chemical Society, pp 165–191.

Stuper AJ, Brugger WE, Jurs PC (1979): "Computer-assisted Studies of Chemical Structure and

Biological Function." New York: Wiley.

Taft RW (1952): Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. J Am Chem Soc 74:2729–2732.

Taft RW (1953): The general nature of the proportionality of polar effects of substituent groups in organic chemistry. J Am Chem Soc 75:4231–4238.

Taft RW (1960): Sigma values from reactivities. J Phys Chem 64:1805–1815.

Taft RW, Ehrenson S, Lewis IC, Glick RE (1959): Evaluation of resonance effects on reactivity by application of the linear inductive energy relationship VI. Concerning the effects of polarization and conjugation on the mesomeric order. J Am Chem Soc 81:5352–5361.

Taft RW, Lewis IC (1958): The general applicability of a fixed scale of inductive effects II. Inductive effects of dipolar substituents in the reactivities of m- and p-substituted derivatives of benzene. J Am Chem Soc 80:2436–2443.

Taft RW, Lewis IC (1959a): Evaluation of resonance effects in reactivity by application of the linear inductive energy relationship V. Concerining a σ_R scale of resonance effects. J Am Chem Soc 81:5343–5352.

Taft RW, Lewis IC (1959b): Evaluation of resonance effects on reactivity by application of the linear inductive energy relationship IV. Hyperconjugation effects of para-alkyl groups. Tetrahedron 5:210–232.

Takemura KH, Cameron MD, Newman MS (1953): Studies on carcinogenic hydrocarbons: Dissociation constants and free-energies of formation of complexes with 2,4,7-trinitrofluorenone. J Am Chem Soc 75:3280–3282.

Tinker J (1981): A computerized structure-activity correlation program for relating bacterial mutagenesis. J Comp Chem 2:231–243.

Topliss JG, Edwards RP (1979): Chance factors in studies of quantitative-structure activity relationships. J Med Chem 22:1238–1244.

Tou JT, Gonzalez RC (1975): "Pattern Recognition Principles." Reading MA: Addison-Wesley.

Tyron RC, Bailey DE (1970): "Cluster Analysis." New York: McGraw Hill.

Verloop A, Hoogenstraaten W, Tipker J (1976): Development and application of new steric parameters in drug design. In Ariens EJ (ed): "Drug Design, Vol 7." New York: Academic Press, pp 165–207.

Waters MD (1979): The gene-tox program. In Hsie AW, O'Neill JP, McElheny VK (eds): "Mammalian Cell Mutagenesis: The Maturation of Tests Systems." Banbury report 2. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, pp 449–467.

Waters MD, Auletta A (1981): The gene-tox program: Genetic activity evaluation. J Chem Inf Comput Sci 21:35–38.

Waters MD, Sandhu SS, Simmon VF, Mortelmans KE, Mitchell A, Jorgenson TA, Jones DCL, Valencia R, Garrett NE (1982): Study of pesticide genotoxicity. In Fleck, RA, Holiaender A (eds): "Genetic Toxicology." New York: Plenum Publishing Co., pp. 275–326.

Weissberg S (1980): "Applied Linear Regression." New York: John Wiley and Sons.

Wepster BM, Hoefnagel AJ, Monshouwer JC, Snorn E (1973): Substituent effects III. Dissociation constants of β-arylpropionic acids, β-arylisovaleric acids, N-arylglycines, aryloxyacetic acids, N-aryl-β-alanines and some related systems. J Am Chem Soc 95:5350–5356.

Wepster BM, van Bekkum H, Verkade PE (1959): Simple reevaluation of the Hammet ρ-σ relation. Rec Trav Chim Pays-Bas 78:815–850.

Wishnok JS, Archer MC (1976): Structure-activity relationships in nitrosoamine carcinogenesis. Brit J Cancer 33:307–311.

Wishnok JS, Archer MC, Edelman AS, Rand WM (1978): Nitrosamine carcinogenicity: A quantitative Hansch-Taft structure-activity relationship. Chem Biol Interact 20:43–54.

Wold S, Dunn WJ III (1983): Multivariate quantitative structure-activity relationships (QSAR): Conditions for their applicability. J Chem Inf Comput Sci 23:6–13.

Wold S, Sjostrom M (1977): SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In Kowalski BR (ed): "Chemometrics: Theory and Application, Symposium Series 52." Washington, DC: American Chemical Society, pp 243–282.

Yalkowsky SH, Flynn GL (1973): Transport of alkyl homologs across synthetic and biological membranes: A new model for chain length-activity relationships. J Pharm Sci 62:210–217.

Yih TD, van Rossum JM (1977a): K_i values of some homologous series of barbiturates and the relationship with the lipophilicity metabolic clearance. Biochem Pharmacol 26:2117–2120.

Yih TD, van Rossum JM (1977b): Pharmokinetics of some homologous series of barbiturates in the intact rat and the perfused rat liver. J Pharmacol Exp Ther 203:184–192.

Yuan M, Jurs PC (1980): Computer-assisted structure-activity studies of chemical carcinogens: A polycyclic aromatic hydrocarbon data set. Tox Appl Pharmacol 52:294–312.

Yuta K, Jurs PC (1981): Computer-assisted structure-activity studies of chemical carcinogens. Aromatic amines. J Med Chem 24:241–251.

Zander GS, Stuper AJ, Jurs PC (1975): Non-parametric features selection in pattern recognition applied to chemical problems. Anal Chem 47:1085–1093.

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.